IMAGE ANALYSIS FOR SPINE SURGERY:
DATA-DRIVEN DETECTION OF SPINE INSTRUMENTATION
&
AUTOMATIC ANALYSIS OF GLOBAL SPINAL ALIGNMENT

by

Sophia Anna Doerr

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Master's of Science and Engineering

Baltimore, Maryland

May 2020

Abstract

Spine surgery is a therapeutic modality for treatment of spine disorders, including spinal deformity, degeneration, and trauma. Such procedures benefit from accurate localization of surgical targets, precise delivery of instrumentation, and reliable validation of surgical objectives – for example, confirming that the surgical implants are delivered as planned and desired changes to the global spinal alignment (GSA) are achieved. Recent advances in surgical navigation have helped to improve the accuracy and precision of spine surgery, including intraoperative imaging integrated with real-time tracking and surgical robotics. This thesis aims to develop two methods for improved image-guided surgery using image analytic techniques. The first provides a means for automatic detection of pedicle screws in intraoperative radiographs – for example, to streamline intraoperative assessment of implant placement. The algorithm achieves a precision and recall of 0.89 and 0.91, respectively, with localization accuracy within ~10 mm. The second develops two algorithms for automatic assessment of GSA in computed tomography (CT) or cone-beam CT (CBCT) images, providing a means to quantify changes in spinal curvature and reduce the variability in GSA measurement associated with manual methods. The algorithms demonstrate GSA estimates with 93.8% of measurements within a 95% confidence interval of manually defined truth. Such methods support the goals of safe, effective spine surgery and provide a means for more quantitative intraoperative quality assurance. In turn, the ability to quantitatively assess instrument placement and changes in GSA could represent important elements of retrospective analysis of large image datasets, improved clinical decision support, and improved patient outcomes.

Primary Reader and Advisor: Jeffrey H. Siewerdsen, PhD

Secondary Reader: Ali Uneri, PhD

Acknowledgements

I would like to acknowledge the opportunity that I have been given to pursue research at an institution such as Johns Hopkins University that is so important in the history of medical advancement and is a model of collaboration between medicine and engineering. True progress in biomedical engineering occurs with clear identification of clinical need. I am grateful for the opportunity to have worked closely with clinical experts on such needs, and I thank Dr. Jeffrey Siewerdsen for illuminating that need through my research. Jeff's strong insight on the interface between medicine, engineering, and industry, as well as his excitement for impactful engineering has taught me a great deal about the need for biomedical engineering in medicine. Through this process, I have garnered a passion to pursue clinical research with hope in the future to think not only of technological advances, but also of system engineering problems at every level of healthcare organizations.

More than anything, I would also like to thank the brilliant clinical minds residing at Johns Hopkins University, with whom I have had the opportunity to communicate or collaborate, such as Dr. Nick Theodore, Dr. Gina Adrales, Dr. William Anderson, and more. The insightful conversations we shared have guided me to understand more about their respective surgical fields and the challenges they face every day.

I am very grateful to the individuals in the I-STAR Lab at Johns Hopkins University with whom I have collaborated, discussed, shared ideas, and gained inspiration to move ideas forward. In particular, the postdoctoral fellows and research scientists at I-STAR who have shared their wealth

of experience with me include Dr. Ali Uneri, Dr. Tharindu De Silva, and Dr. Craig Jones – to whom I am very grateful. To all the wonderful people from whom I have learned at Johns Hopkins, I am grateful for the impact you have had on my life.

Dedication

This thesis is dedicated to the pursuit of knowledge and good of humanity. Although a small contribution, I hope to continue to better the world through the knowledge and experience this work has taught me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The prevalence of spinal disorders – including spinal deformity, degeneration, and trauma – presents a large burden on societal health and the healthcare system. Slow progression of some spinal disorders combined with a lack of preventative health measures and delays in diagnosis and treatment contribute to low and variable levels of treatment efficacy.[1] Spine surgery is an important therapeutic modality for treatment of spinal deformities, degeneration, and trauma, with the total number of spine operations in 2002 exceeding one million.[2] In 2015, there were 199,140 elective lumbar fusions at a cost of around $10 billion.[3] More than 50% of first-time spine surgeries achieve positive (i.e., beneficial) patient outcomes; however, the rate of success in revision surgery decreases successively (30%, 15%, and 5% with second, third, and fourth surgeries respectively).[2] Precise administration of treatment is imperative to reduce the need for revisions.

Among the contributors to technical failure in spine surgery is targeting of the wrong vertebral level – so-called "wrong-level" surgery, often associated with manual counting errors, with a reported incidence of 1 in 3110 spine surgeries.[4] Anatomical variation and degenerative pathologies can confound target localization even for experienced surgeons, challenging accurate treatment delivery.

Another contributor to technical failure is malplacement of surgical instrumentation – for example, failure to deliver screws safely or securely within the spinal pedicle, leading to comorbidity and/or implant failure. The rate of misplaced pedicle screws has been shown to range from 14% to 55% using standard insertion techniques not involving navigation assistance.[5]

Image guidance in spine surgery improves visualization of the target, surrounding tissues, and surgical instrumentation, helping to improve the accuracy of pedicle screw placement and reduce the rate of pedicle breach to less than 5%.[5] Surgical navigation using preoperative and/or intraoperative imaging has evolved over the last two decades to assist in precise localization, placement of instrumentation, and promotion of minimally invasive surgical techniques.[6]

In spine surgery, intraoperative x-ray radiography or fluoroscopy is often used for visual verification of patient anatomy and implant placement, although visualization and 3D localization can be confounded by anatomical clutter in projection images and the inability to precisely determine 3D information from visual interpretation of 2D projection views. Surgical navigation based on preoperative CT has shown improved accuracy in pedicle screw placement.[7] Postoperative CT can also be used to validate pedicle screw placement,[8] but acquiring images following wound closure and outside the operating room (OR) introduces workflow difficulties and a challenging clinical decision for costly revision surgery. Intraoperative CT or cone-beam CT (CBCT) can be used to register to preoperative CT and evaluate instrument placement in the OR at the time of treatment. Such systems offer to improve the accuracy, precision, and safety of spine surgery and require streamlined integration with clinical workflow for broad utilization.[5] Although surgical navigation has demonstrated improved levels of precision in implant placement,[7] challenges still remain with respect to clinical outcomes, and improvements in workflow integration and automation in image analysis can promote better assessment of the surgical product. The ability to automatically localize targets and verify implant placement in 2D radiography / fluoroscopy offers major benefits to mainstream use.

Accurate placement of surgical implants in spine surgery correlates with improved patient outcomes, as precise placement provides high pullout strength and prevents pedicle breach.[5] Reliable intraoperative assessment of the surgical product provides the opportunity to revise implant placement prior to leaving the OR and helps to reduce the frequency of costly revision surgery. Current evaluation of screw placement is commonly performed by qualitative assessment of an intraoperative radiograph or – if intraoperative CT or CBCT is available – by assessment of a 3D image acquired at the end of the case.

Other methods have been proposed to quantitatively assess device placement through 3D-2D registration of the preoperative CT to a post-instrumentation, intraoperative radiograph – for example, using the known-component registration (KC-Reg) algorithm.[9] While KC-Reg has demonstrated a high degree of accuracy in 3D localization of surgical instruments from 2D projections, some challenges remain in order for it to meet efficient workflow requirements in surgery. These include the need to *initialize* the registration with

a reasonable estimate of device location, orientation, and model (i.e., type of implant). Clinical translation of an algorithm such as KC-Reg would greatly benefit from automatic detection of instrumentation in intraoperative radiographs.

In addition to safe, accurate, and precise delivery of surgical instrumentation, the success of spine surgery also relies on the changes imparted to the anatomy. For example, surgical correction of spinal deformity often aims to impart a change in spinal curvature, which is typically quantified in terms of various metrics of global spinal alignment (GSA). GSA metrics are pertinent to clinical decision support in treating spinal deformity and disability and correlate well with patient outcomes.[10–13] Assessing GSA in the OR is challenged by a number of factors, including difficulty in visualizing spinal curvature on the operating table and understanding its relationship to measurements under weight-bearing conditions. Current standard metrics for evaluation of spinal morphology are manually defined in post-operative radiographs. The ability to accurately measure GSA from images acquired in the OR could provide a valuable means of verification and quality assurance (QA) and help improve clinical outcomes. Manual methods lack reproducibility due to high reader variability.[14] Automatic measurement of GSA in intraoperative radiographs could provide clinical decision support and provide quantitative assessment of changes in GSA.

Image analytics, including quantitative measures of implant placement and changes in spinal curvature, have the potential to improve accuracy and patient outcomes in spine surgery. A recent study in a large database of spine surgery scans demonstrated that image features – as opposed to demographic features – had the strongest correlation to patient outcomes, in which local curvatures about T12-L1 and L4-L5 vertebrae were the most important features in predicting 12-month mJOA outcome.[15] Quantitative metrics extending from image-derived features are already used in other surgical fields, such as radiation oncology, in which treatments are delivered through optimization of radiation dose distribution to the treatment volume.[16] Considering the challenges faced in spine surgery outcomes, metrics derived objectively and consistently from readily available image data could help to improve patient outcomes and reduce the rate of failed back surgery.

This thesis aims to develop image analysis techniques that accurately and automatically (or semi-automatically) localize surgical instrumentation (viz., spinal pedicle screws) and changes in patient anatomy (viz., measures of GSA) in intraoperative images. In Chapter 2, a method is developed for automatic detection and localization of pedicle screws in intraoperative radiographs using artificial neural networks. In Chapter 3, methods are developed to analyze metrics of GSA from CT or CBCT. Together, these methods could provide a basis for improved intraoperative QA, more quantitative clinical decision support, and a basis for retrospective analysis of large image datasets in support of data-intensive analytics and prediction of clinical outcomes.

Chapter 2: Data-Driven Detection of Spine Surgery Instrumentation in Intraoperative Images

1 Introduction

Accurate placement of spinal instrumentation is essential to mitigating complications (e.g., breach of the spinal canal),[17] ensuring stable fixation (high pullout strength), and improving long-term patient outcomes. Intraoperative checks on device placement can be performed using radiography/fluoroscopy, but assessment is largely qualitative, and poor image quality can confound reliable evaluation of implants in relation to 3D patient anatomy. Previous work established a method for quantitative verification of implants using 3D-2D image registration,[9] combining preoperative CT of the patient registration and analysis based on the known designs of surgical implants to achieve accurate 3D localization.

Despite the demonstrated accuracy of such model-based registration approaches, performance can be challenged by a number of factors, including: (1) limited capture range; (2) model validity (i.e., ensuring that the correct screw model is used); and (3) stringent runtime requirements of the intraoperative workflow. Data-driven approaches, such as deep learning, have shown great potential for complex image processing tasks that rely on both local and global contextual information.[18–20] Convolutional neural networks (CNNs), for example, have been proposed to detect the geometric pose of surgical screws in 2D and 3D radiographic images,[21] with varying degrees of geometric accuracy and sensitivity to training data.

In this work, a deep CNN was developed to count, localize, and classify pedicle screws as an initialization for a subsequent model-based known-component registration (KC-Reg).[9] The combined approach brings together the speed and global support of data-driven approaches with the accuracy of model-based techniques. Strong initialization from the CNN is hypothesized to enable KC-Reg to run faster, with fewer iterations, and less susceptibility to local minima. The combined registration system offers to improve workflow for a number of algorithms that have been developed using KC-Reg for guiding instrument placement, verifying implant placement, and reducing metal artifacts.[22,23]

2 Methods

2.1 Model-Based 3D-2D Registration

The proposed solution is intended to initialize the KC-Reg algorithm for 3D-2D image registration, which uses prior knowledge of surgical instrument shape and material type.[9] The registration solves for the 3D pose of the screws using an evolutionary optimizer (CMA-ES) to iteratively correlate image features (gradients) between the radiographic views and projections simulated at estimated poses. Previous work has demonstrated application of this method on complex models with articulating pieces (e.g., polyaxial tulip heads), simple parametric models, and models that allow deformation of the implant (e.g., flexible K-wires and stents). KC-Reg has also been incorporated in other algorithms – for example, the "known-component metal artifact reduction" (KC-MAR) algorithm[22] and the "known-component reconstruction" (KC-Recon) algorithm[23], where the registration process provides sub-pixel segmentation of device boundaries prior to (or in joint optimization with) image reconstruction. Accurate delineation of metal implants was shown to reduce streak and blooming artifacts and permit clear visualization of anatomical structures at the implant boundary.

Figure 2.1 demonstrates the proposed workflow, with a component initializer that counts, localizes, and classifies component models from input intraoperative radiographs to initialize the "known component" parameters for KC-Reg. The data-driven initialization stage provides initial component localizations to constrain the optimization search space, avoid local minima, and reduce registration runtime. Model selection leverages the estimated bounding box size in multiple views as well as detected screw length and width. Screw localization is the subject of work detailed below, and screw classification ("model selection" in Figure 2.1) is the subject of future work – e.g., identifying a particular screw type based on the bounding box size observed in multiple views as well as the detected screw length and width.

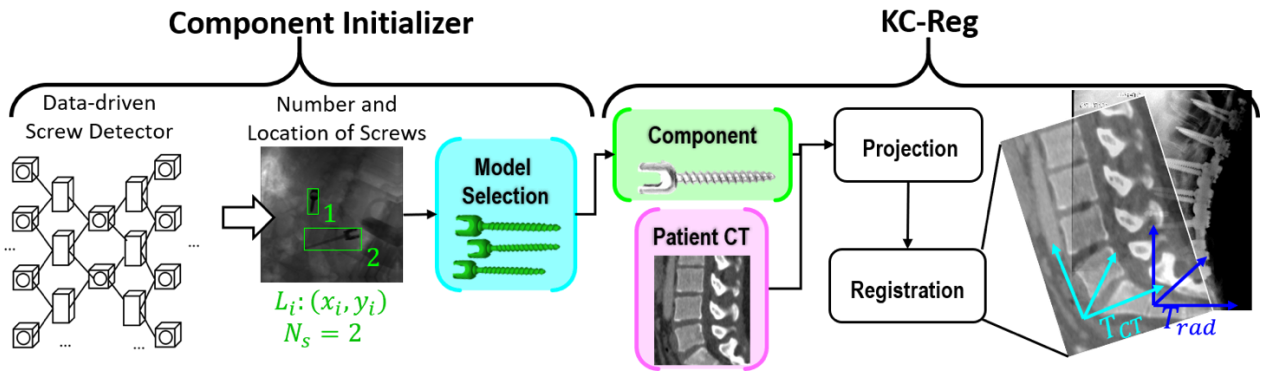

Figure 2.1. Algorithm workflow for intraoperative 3D-2D registration of preoperative CT to post-instrumented intraoperative radiographs. (a) The component initializer is composed of data-driven detection of screws – yielding the number (N_s) and location (L_i) of screws – in radiographs and model selection. (b) KC-Reg uses knowledge of initial component location and model type to perform accurate 3D-2D registration.

2.2 Data-Driven Detection of Pedicle Screws in Intraoperative Radiographs

Object detection networks have demonstrated strong performance in detection-related tasks and have been proposed for use in medical imaging. Faster R-CNN, for example, performs particularly well in a diverse range of datasets with high accuracy and precision.[24] The network architecture for this work uses Faster R-CNN with VGG16[25] for feature extraction. Faster R-CNN consists of three distinct segments: the backbone feature extractor; a region proposal network (RPN); and a detection network (RCNN). The feature extractor inputs salient image map features to both the RPN and RCNN. These feature maps are extracted through a series of convolutional layers in the backbone network. The RPN then determines which blocks of a 16×16 grid of the input image are most likely to include objects (by proposing anchor boxes for each block) and determines respective transformation coordinates to refine anchor box proposals to more closely fit objects. The RCNN performs crop pooling to output a 1D feature vector for each ROI, performs classification (background vs. screw) on the ROIs, and outputs bounding box coordinates by class.

Input training data consisted of projection images and ground truth bounding box labels. Data generation combined real patient CT images and screw models as described in greater detail in Section 2.3. Bounding box labels were defined as a vector of minimum and maximum x and y coordinates. Coordinates were

automatically obtained from the simulation process illustrated in Figure 2.2, where output screw masks give the extent of the object in the image.

The network outputs bounding box coordinate labels, from which the number of screws and 2D locations are determined. Post-processing can be used to further refine screw localization. In ongoing work, the 3D pose of each screw is determined from detection of the same screw in multiple radiographs. Screw location, pose, and number are taken as input to initialize KC-Reg in order to relate the 3D position as determined from intraoperative radiographs to a preoperative CT.

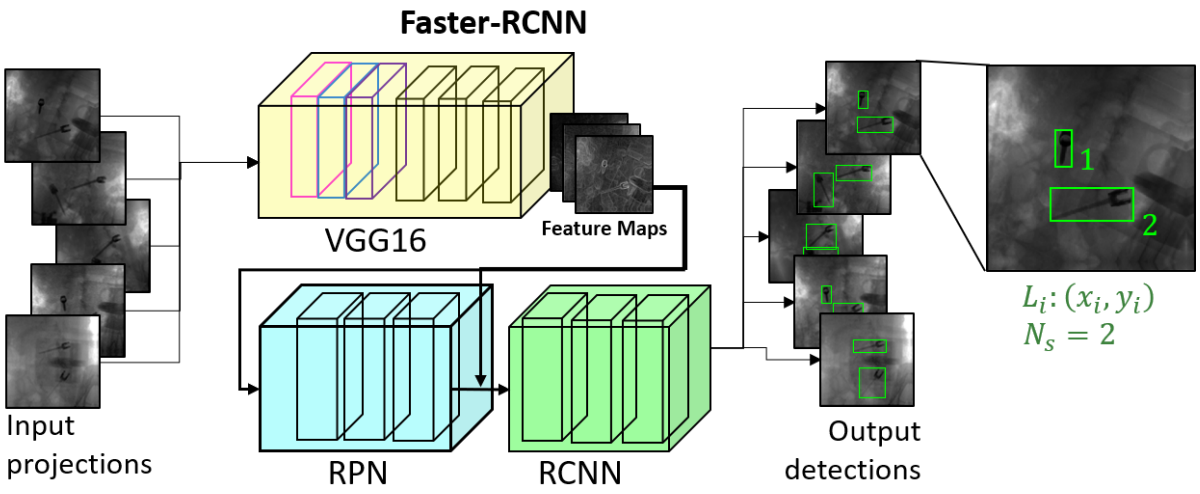

Figure 2.2. Convolutional neural network architecture for screw detection in radiographs. The network achieves fast 2D localization of screws as initialization to subsequent model-based image registration and reconstruction approaches.

2.3 Training Dataset of Realistic Surgical Instrumentation

A clinical study was performed with Institutional Review Board (IRB) approval in which 17 subjects undergoing spine surgery were imaged under informed consent. 3D cone-beam CT (CBCT) scans were acquired of each subject using the O-arm (Medtronic, Littleton MA) prior to and following instrument placement, recording the screw models delivered in each case. Standard clinical techniques were used (120 kV with 186–596 mAs), yielding 745 projections of 1024×384 pixels at 0.388×0.766 mm^2 spacing (at 2×4 detector binning) for each scan.

The training dataset was generated from the projection datasets acquired with each CBCT scan combined with the design specifications (CAD models) of Solera spine instrumentation (Medtronic, Littleton MA) – specifically, 30 types of pedicle screws. The image simulation pipeline depicted in Figure 2.3 was constructed to combine real (pre-instrumentation) projection images with screws placed virtually in the 2D projections, yielding 13,000 annotated radiographic images.

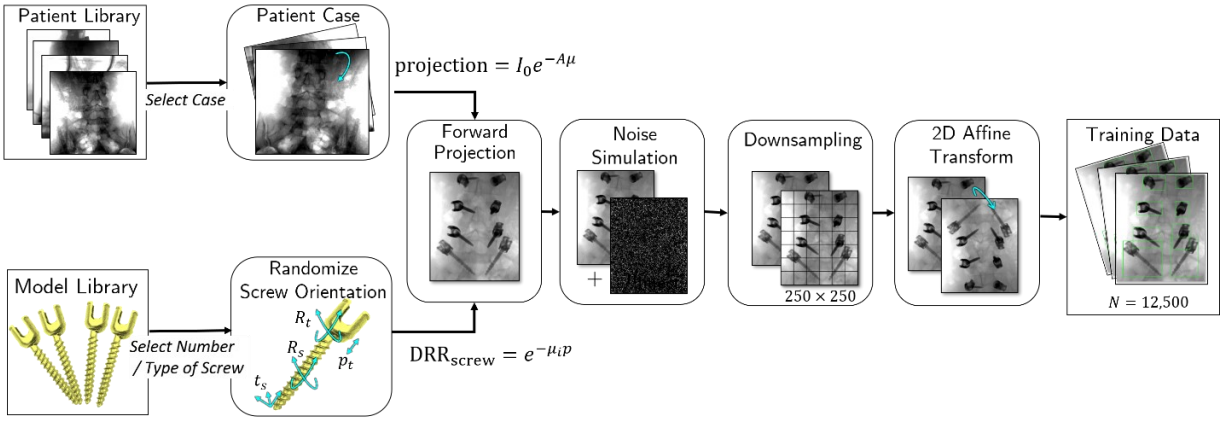

Figure 2.3. Generation and augmentation of the training dataset. Augmentations of the clinical images included affine transformations (without skew), downsampling, and noise simulation. The initial pipeline was limited to 5 screw types. While the data generation pipeline used images from 17 subjects, each step in the process introduced an increasing number of random elements to augment the number of training images.

For each subject, a 2D projection image was randomly sampled from the raw projections of the pre-instrumented CBCT scan (potentially containing other tools, such as retractors) to realize a highly realistic, complex radiographic scene. View selection included projection angles ranging from AP to LAT, including oblique views. The images were cropped to square format (from the 40×30 cm^2 detector size) and downsampled using cubic interpolation to 250×250 px at isotropic spacing of 1.2×1.2 mm^2. For simulating the screws, combinations of screw pairs were randomly sampled from the model library. Each screw was oriented according to previously defined transpedicle trajectories at a particular vertebral level. The images were further augmented by imparting a random translation (Gaussian distribution with σ = 10 mm) and rotation (σ = 7°) to each screw trajectory, with additional rotation (σ = 15°) for the polyaxial tulip head. The screws were then projected onto (i.e., added to) the downsampled projections. Random noise approximating quantum and electronic noise was added to simulate various levels of imaging dose[26] and to

allow the network to learn salient features of the screws (and ignore quantum noise). Finally, the training set was augmented by translation (−20, 20 px) and rotation (−15°, 15°) of the images.

Detection labels were computed directly from the projection of the screws into the 2D radiograph, where extent in x and y were known. Subsequent bounding box coordinates and input projection images were associated for training. A total of 12,500 images was used for initial network training, beginning in the current work with simulations based on images of a single patient (i.e., many projection views of common underlying 3D anatomy). Ongoing work will include a diversity of patient cases in the training set. As shown below, training based on images with anatomical background derived from one patient presented a test of network generalizability to other patient cases and the network's ability to ignore background anatomy in the screw detection task.

2.4 Model Training

Network training was completed on the training dataset of 12,500 images described in Section 2.3 for 500 epochs on a GeForce RTX (Nvidia, Santa Clara CA). A batch size of 128 images per iteration was used with a training / validation split of 90% / 10%. The loss function consisted of a sum of categorical cross entropy and a smoothing L1 regression loss for the bounding box coordinates. An Adam optimizer with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used for training. Optimal learning rate was determined through a series of learning rate sweeps from 10^{-2} to 10^{-10}. The lowest loss (and highest detection accuracy) was observed between 10^{-6} and 10^{-8}, so an initial learning rate of 10^{-6} with a learning rate decay of 0.1 after 250 epochs was chosen.

2.5 Model Testing

Testing was performed for training on a single patient dataset and on a multiple patient dataset. Testing for the single patient training consisted of two datasets The first dataset consisted of 2,000 images generated from the same patient, but with random instantiations of screws, screw poses, projection views, noise, and affine transformations as detailed in Section 2.3 that differed from the initial training dataset. The second dataset consisted of 2,000 images generated from five patients with similar random instantiations of

parameters as the first dataset. The performance on the first dataset (referred to below as the "same-patient" dataset) demonstrates the ability of the network to generalize due to screw types, poses, viewing angle, and noise, whereas the second dataset (referred to below as the "many-patient" dataset) demonstrates the robustness of the network to varying background anatomy.

Multiple patient training was evaluated with two validation folds. The multiple patient dataset consisted of six cases. To test on data that hadn't been seen before, the first test consisted of training on four of the patients and testing on two of the patients (training cases: 2, 4, 5, 6, testing cases: 1, 3) and the second test consisted of training on a different set of four and testing on a different set of two patients (training cases: 2, 4, 5, 6, testing cases: 2, 4). Two separate testing folds were chosen to assess generalizability through independence of the training and testing data.

The accuracy of screw detection was evaluated in terms of the precision:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

as well as the recall:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. A total of 500 images were sampled from each test dataset to compute these measures of detection accuracy.

The accuracy of screw localization was evaluated in terms of the intersection-over-union:

$$\text{IOU} = \frac{\text{A} \cap \text{B}}{\text{A} \cup \text{B}} = \frac{\text{TP}_\text{i}}{\text{TP}_\text{i} + \text{FP}_\text{i} + \text{FN}_\text{i}} \tag{2.3}$$

where region A represents the predicted bounding box, region B represents the ground truth bounding box, TP_i is pixelwise true positives for an object, FP_i is pixelwise false positives for an object, and FN_i is

pixelwise false negatives for an object. Localization accuracy was also evaluated in terms of the difference (distance) between predicted and true bounding box coordinates.

Sample images from training and testing datasets are shown in Figure 2.4, illustrating variations in screw pose, noise level, and projection view angle.

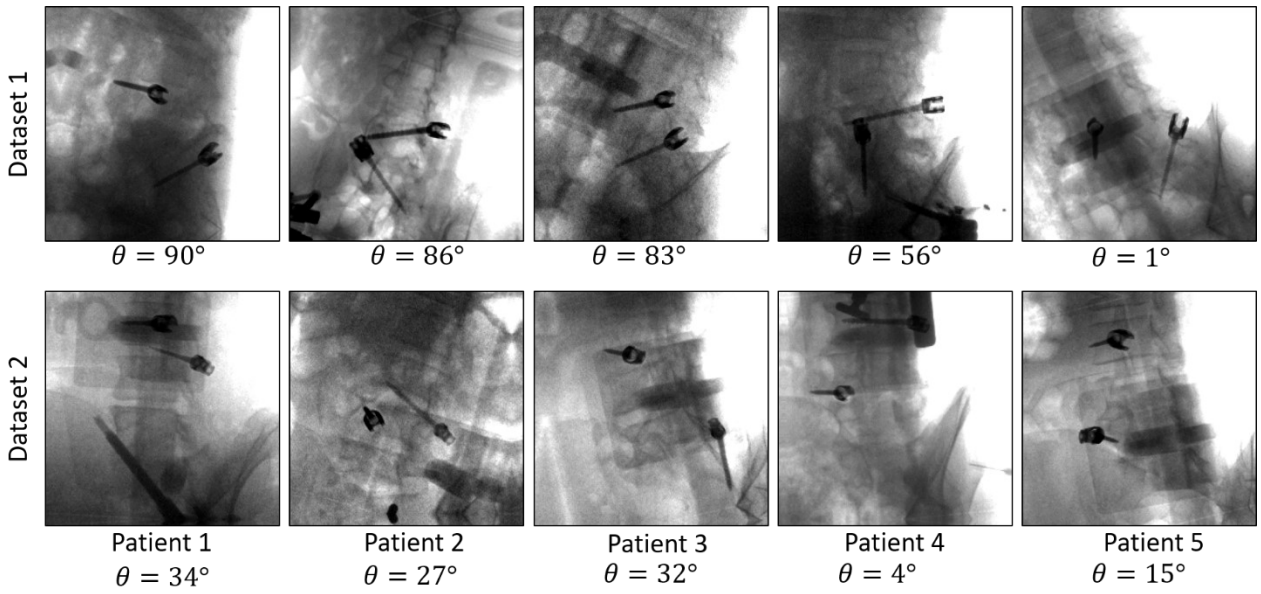

Figure 2.4. Example radiographs with instances of pedicle screws to be detected. (a) Example images from the training set, which consisted of 12,000 projections at various view angles, each with two pedicle screws. (b) Example images from dataset 2, consisting of projections from five patients at various view angles, each with two pedicle screws.

3 Results

3.1 Model Testing

Predictions were generated for single-patient training on test dataset 1 (same-patient) and dataset 2 (many-patient) with 2,000 images each. Sample detections for test dataset 1 are depicted in Figure 2.5 and appear reasonably close to ground truth labels. This suggests that the network performed reasonably well for a wide variety of screw poses, noise levels, view angles, and affine transformations of the data.

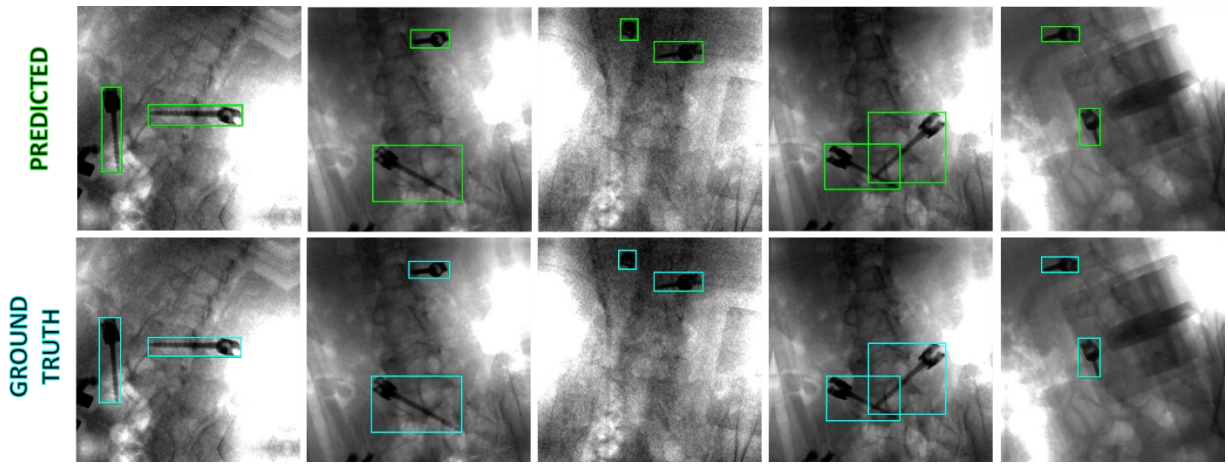

Figure 2.5. Example network detections for test dataset 1. Predicted bounding boxes (green) closely resemble the ground truth (cyan).

Sample detections for five patients in test dataset 2 are depicted in Figure 2.6. Detections again appear to be in line with ground truth labels, despite not having seen such patient anatomy during training. In the sample detection for patient 4, there is also another surgical instrument (other than a screw) in the background of the image, for which predictions are still accurate. The results suggest that the network generalized fairly well in learning to ignore background anatomy and other surgical instrumentation and was able to discern salient image features specifically associated with the pedicle screws. For multiple-patient training, predictions are shown in Figure 2.7, with predictions appearing close to ground truth. Some slight deviations are noticed in the edges of the bounding boxes, though barely noticeable. The detection ability of the network suggests that the varied background anatomy did not affect the ability of the network to learn the appearance of screws in the images. It is worth noting that the multiple-patient training data was log-normalized, so the screws appear lighter in these images. Log normalization was used to better distribute the range of image intensities used by the network.

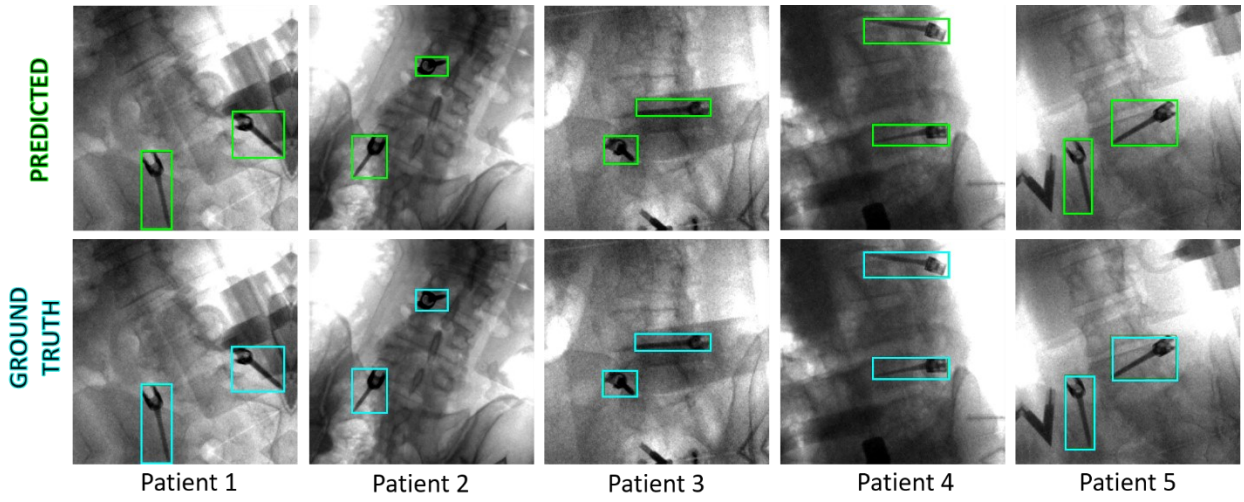

Figure 2.6. Sample network detections for test dataset 2. Predicted bounding boxes (green) appear to closely resemble the ground truth (cyan).

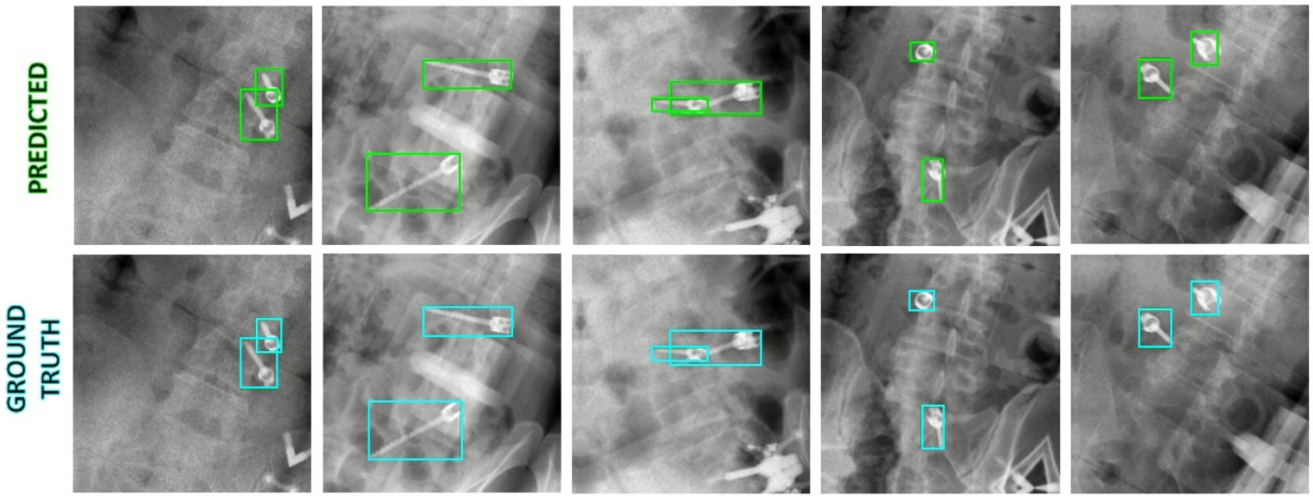

Figure 2.7. Sample network detections for Multiple patient training. Predicted bounding boxes (green) appear to closely resemble the ground truth (cyan).

Detection accuracy was computed for single-patient training on test datasets 1 and 2, determining the ability of the network to accurately count the number of screws in each image. Figure 2.8 shows the true-positive, false-positive, and false-negative detections for both test datasets. Both test datasets exhibited a high true-positive rate and a low false-positive rate. Notably, the performance of dataset 1 exceeded that of test dataset 2, suggesting some residual advantage for the same-patient training case. Precision and recall are shown for multiple patient training in Figure 2.9, where the data show an average precision of ~92% and recall of

~89%. Neither dataset exhibits a considerable number of false or missed detections. The precision and recall are shown for single-patient and multiple-patient training in Table 2.1.

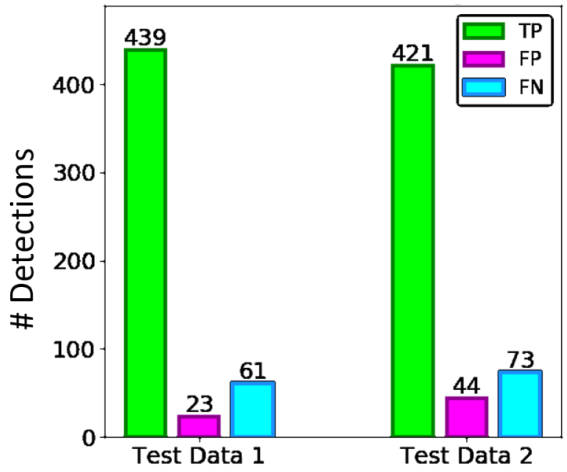

Figure 2.8. Performance of screw detection for single-patient training: true-positive, false-positive, and false-negative detections for test datasets 1 and 2.

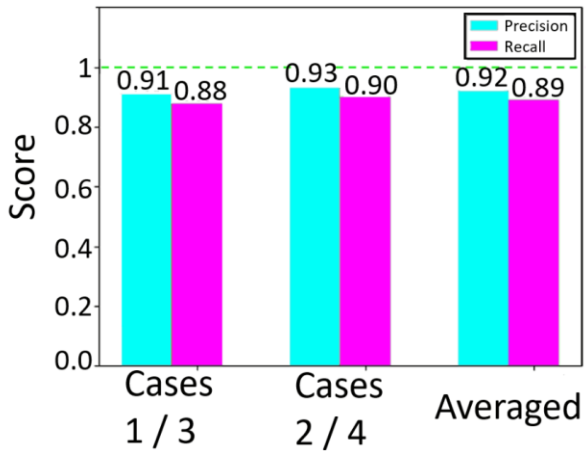

Figure 2.9. Performance of screw detection for multiple-patient training: precision and recall detections for cross validation sets with testing on cases 1/3, cases 2/4, and averaged.

			Precision	Recall
Single	**Patient**	**Test dataset 1**	95.0%	87.8%
		Test dataset 2	90.5%	85.2%
		Overall (Mean)	**92.8%**	**86.5%**
Multiple	**Patient**	**Cases 1 / 3**	91.0%	87.6%
		Cases 2 / 4	93.2%	89.9%
		Overall (Mean)	**92.1%**	**88.8%**

Table 2.1. Performance of screw detection: precision and recall.

3.2 Accuracy of Localization

The spatial accuracy of screw localization was evaluated in comparison to ground-truth bounding boxes. The localization accuracy is quantified in terms of IOU in Figure 2.10, showing the fraction of the predicted bounding box area that is shared with ground truth. The overall median IOU for single patient training on both datasets was 0.90 (with interquartile range, IQR ~0.77 – 1.0). The median IOU was 0.91 (with IQR ~0.8 – 1.0) for test dataset 1, and the median IOU for dataset 2 was 0.87 (with IQR ~0.73 – 1.0). Accordingly, both test datasets appear to perform well in localization as measured by IOU, with the same-patient dataset performing slightly higher (p-value = 0.08) possibly due to residual influence of the learned background anatomy. Multiple-patient training IOU in Figure 2.11 shows a median IOU of 0.91 for two pooled cross validation datasets. It appears that training on multiple patients (in this case, 4) overcame the differences in background anatomy between test dataset 1 and 2 in the single patient training.

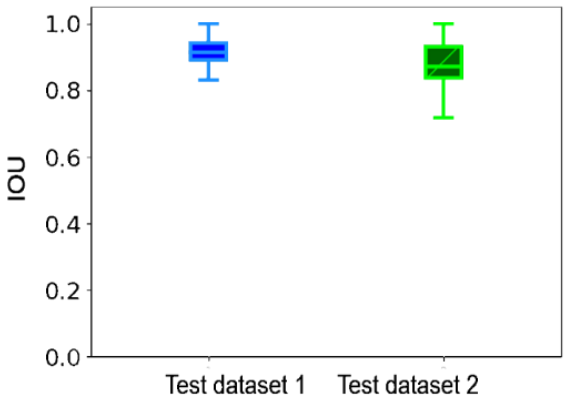

Figure 2.10. Localization accuracy of screw detections for single-patient training: IOU of predicted and ground truth bounding boxes.

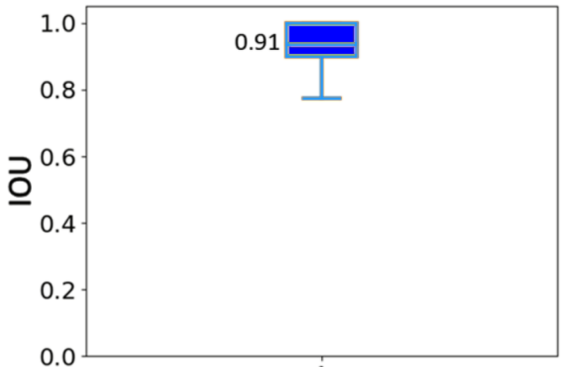

Figure 2.11. Localization accuracy of screw detections for multiple-patient training: IOU of predicted and ground truth bounding boxes.

The spatial accuracy of screw localization was further quantified in terms of the mean difference in (distance between) bounding box coordinates. Note that the coordinate difference metric captures errors associated with a case in which the detected bounding box is centered on a screw but its size differs from the true size of the screw. The mean of the difference of ground truth and predicted bounding box coordinates for single patient training is shown in Figure 2.12. The median bounding box coordinate differences were 1.2 mm for test dataset 1 and 0.4 mm for test dataset 2. Test dataset 2 appears to have a slightly broader spread of localization error than test dataset 1. Overall, the network was able to localize screws within ~10–13 mm of ground truth screw (including outliers), which is within the typical capture range of KC-Reg. For multiple patient training, shown in Figure 2.13, localization error seems to be within ~10 mm, again within the typical capture range of KC-Reg.

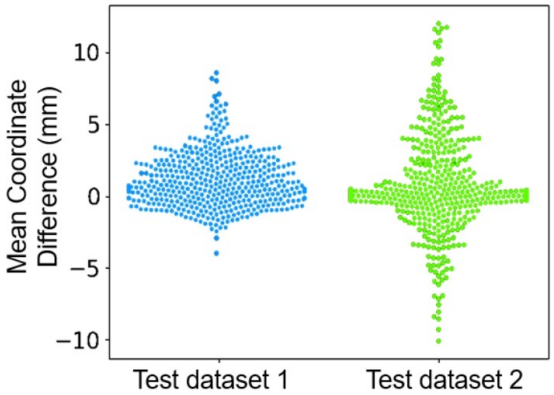

Figure 2.12. Localization accuracy of screw detections for single-patient training: Distance between predicted and ground truth bounding box coordinates.

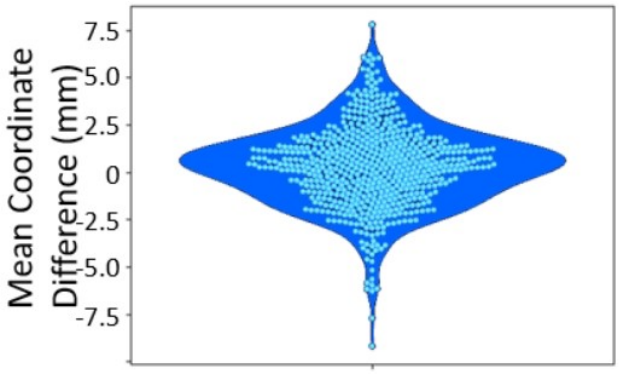

Figure 2.13. Localization accuracy of screw detections for multiple-patient training: Distance between predicted and ground truth bounding box coordinates.

4 Discussion and Conclusions

A data-driven method was presented using a deep CNN to initialize pose estimates for subsequent model-based registration and reconstruction of intraoperative images containing surgical instrumentation (viz., spine screws). The study demonstrates a potentially streamlined approach to initializing model-based registration approaches, such as KC-Reg, to better integrate with intraoperative surgical workflow. To this end, the CNN approach provides initialization that could obviate the need for manual user input, help to avoid local minima, and reduce runtime by limiting the search space. The network demonstrated high accuracy in single patient and multiple patient screw detection – counting screws with fairly high precision (~92.6%, 92.1%) and recall (~86.5%, 88.8%). Localization performance demonstrated median IOU ~0.90 and median bounding box coordinate difference within ~1-2 mm (and within ~10–13 mm including outliers). This level of initialization accuracy was well within the typical capture range of KC-Reg.[9]

Among the limitations in the initial development of the network is the scope of training data, which consisted of projection data from CBCT scans of a single patient with scan geometry restricted to that of the O-arm imaging system. The screw models projected in both the training and test datasets in the current work were sampled from a library of Solera spine screws, and the generalizability to other designs of pedicle screws is unknown at the current time. Moreover, the current work involved training data presenting only two screws. The approach likely extends to more complex scenes with a higher density of implants – to be investigated in future work. Furthermore, training on a more broadly varied set of projection data is anticipated to yield greater network generalizability, including training data with increased variability in

the number of screws, patient anatomy, and imaging protocol (e.g., energy and dose). Ongoing work extends the network detection to integrating multiple views of the same scene.

Overall, the data-driven approach to screw detection in intraoperative radiographs appears to provide a useful means of initializing model-based registration (KC-Reg)[9] and reconstruction (KC-MAR and KC-Recon)[22,23] approaches. Such initialization is important to the stability, robustness, and runtime of such algorithms and likely an important step to their realization in routine clinical use. Given the promising initial performance, one might envision future development in which the data-driven approach in itself may provide localization accuracy sufficient for the 3D-2D registration – a possibility that forms the subject of future work.

Chapter 3: Automatic Analysis of Global Spinal Alignment

1 Introduction

Quantitative measurements of global spinal alignment (GSA) are frequently used to evaluate spinal deformity and characterize the progression of spinal disability.[10–12] GSA metrics are conventionally measured by manual annotation of anatomical landmarks (e.g., vertebral endplates) in radiographic images and have been shown to correlate with spine surgery outcomes in treatment of both deformity and degenerative disease.[10–13,27] For example, courses of treatment for scoliosis and kyphosis are frequently chosen based on progression of GSA measures beyond pre-determined thresholds. (E.g., normal thoracic kyphosis falls within a range 20–40°).[28] Retrospective analysis[29] of patients undergoing degenerative lumbar spinal fusion suggests that GSA assessment is pertinent to lumbar fusion as well, with abnormal ranges of spinopelvic GSA parameters (pelvic incidence – lumbar lordosis mismatch, PI–LL > 11°) indicating a 75% positive predictive value for adjacent segment failure.

The conventional approach to measurement of GSA is based on manual identification of pertinent anatomical landmarks in radiography. However, variability in manual annotation is high. As summarized in Table 3.1, intra-reader intraclass correlation coefficient (ICC) has been reported to range from 0.40 to 0.96 (poor to excellent).[30] Inter-reader ICC shows similar findings (ICC ranging from 0.20 to 0.86) and confirms expectations that variability between readers is worse than variability within one reader. Such wide variability implies important dependence on study controls and conditions and suggests inherent variability in GSA measurement based on visual assessment of endplate angles. As reported by Carman et al.,[31] change in GSA must exceed 11° to be of clinical utility and not be solely attributable to measurement error. Challenges to manual measurement include radiographic image noise,[14] angulated projection of the endplate plateau,[14,30,32] differences in plateau visualization due to kyphosis and lordosis, patient obesity, and degeneration of bone quality.[30] The lack of reproducibility in manual measurement has led some to question the usefulness of measurements of sagittal alignment from radiographs.[14]

Inter-Reader ICC				Intra-Reader ICC				Reference
Thoracic	**Lumbar**	**Pelvic / Global**	**Average**	**Thoracic**	**Lumbar**	**Pelvic / Global**	**Average**	
0.50-0.57	0.76-0.84	-0.04-0.44	**0.29-0.95**	0.36-0.46	0.62-0.74	–	**–**	[33]
–	–	0.50	**0.71**	0.83	0.70	0.77	**0.71**	[32]
>0.80	<0.80	>0.80	**>0.80**	–	>0.80	<0.20-> 0.80	**0.40-0.76**	[14]
0.88	0.90	0.83	**0.81**	–	–	0.86	**0.87**	[34]
–	–	0.83	**0.89**	0.79-0.99	0.84-0.99	–	**–**	[35]
0.82-0.85	0.82-0.93	–	**–**	0.78-0.86	0.87-0.96	0.78-0.86	**0.82-0.96**	[30]

Table 3.1. Summary of reports on inter- and intra-reader variability (ICC) in measurement of spinal alignment.

Computer-assisted technologies – for example, Surgimap (Globus Medical, Audubon PA) and iGA (NuVasive, San Diego CA) – have shown similar or better performance than manual approach in analyzing GSA.[36] However, such approaches still require significant manual input that is subject to variability. Recently, a number of algorithms for automated vertebral labeling and segmentation have emerged that could improve the identification of image features pertinent to GSA to increase efficiency and reduce variability. Such automatic labeling methods include model-based approaches (prior shape modeling,[37] mathematical morphology,[38] regression forests[39]) and deep learning approaches,[18,20,40] some of which have been translated to clinical use – for example, FAST Spine (Siemens Healthineers, Malvern PA).

Automatic determination of GSA metrics would be of benefit not only in diagnostic radiology (e.g., assessment of scoliosis) but also in intraoperative measurement of GSA as a tool for quantitatively evaluating the change in curvature imparted during spinal reduction. As discussed by Ailon et al.,[41] the development of an intraoperative tool for GSA computation requires efficient computation time and a low impact on workflow, further motivating an automatic method. The evolution of minimally invasive, image-guided spine surgery in recent decades raises the opportunity for more quantitative assessment of GSA in the perioperative context. A straightforward method involves use of a surgical tracking system to localize pertinent anatomical features on the patient and assess GSA accordingly. Image-guided spine surgery has also prompted more frequent acquisition of preoperative and postoperative CT – from which changes in

GSA may be determined – and steady adoption of intraoperative CT from which assessment of deformity correction could be performed during the case.[42–44]

Automatic assessment of GSA from preoperative, intraoperative, and postoperative spine CT or radiographs could provide a valuable tool for data-intensive methods that aim to analyze large-scale image datasets for correlation with patient outcomes. For example, recent work by DeSilva et al.[15] shows that GSA (among other features derived from perioperative images) could provide insight on surgical outcomes beyond that of conventional patient demographic data. It also prompts further analysis of the relationship between GSA assessed in supine, prone, and standing orientations.

The work reported in this chapter details automatic measurement of radiographic GSA metrics from vertebral labels identified in spine CT using two distinct approaches. The first approach (referred to as EndSeg) operates by segmentation of vertebral endplates, analogous to conventional manual approach based on measurement of endplate angles. The second approach (referred to as SpNorm) uses the vertebral labels themselves as a surrogate for spinal curvature determined by spline fit. Both operate by projection of CT-derived inputs to 2D radiographic planes for GSA analysis in terms that are consistent with conventionally defined 2D GSA measures. Each method is tested in retrospective analysis of spine CT images and compared to manual definition by expert radiologists.

2 Methods

Metrics of GSA are defined by various Cobb angles as illustrated in Figure 3.1. On a radiograph, a Cobb angle is the angle between lines parallel to the superior endplate of a particular (superior) vertebra and the inferior endplate of a second (inferior) vertebra as presented in a 2D radiographic view. Three methods for GSA measurement are described below and evaluated in terms of reproducibility and accuracy: conventional manual annotation and two novel automatic methods.

	GSA Metric	Definition
Sagittal	C7S1 Sagittal Vertical Axis (SVA)	Horizontal distance from vertical plumb line through C7 to posterior-superior corner of S1 end plate
	Proximal Thoracic Kyphosis (PThK)	Angle subtended by line parallel to superior endplate of T1 and line parallel to inferior endplate of T5.
	Main Thoracic Kyphosis (MThK)	Angle subtended by line parallel to T4/T5 superior and line parallel to inferior endplate of T12.
	Thoracic Kyphosis (ThK)	Angle subtended by line parallel to superior endplate of T1 and line parallel to inferior endplate of T12
	Lumbar Lordosis (LL)	Angle subtended by line parallel to superior endplate of L1 and line parallel to superior endplate of S1
	Pelvic Incidence (PI)	Angle between line from hip axis to midpoint of sacral endplate and line orthogonal to center of sacral endplate
	Pelvic Tilt (PT)	Angle subtended by vertical reference line through hip axis and line from midpoint of sacral endplate to hip axis
	Hip Axis (HA)	Midpoint between approximate centers of both femoral heads
Coronal	Proximal Thoracic Cobb Angle (PThC)	Angle subtended by line parallel to superior endplate of T2 and line parallel to inferior endplate of T4.
	Main Thoracic Cobb Angle (MThC)	Angle subtended by line parallel to superior endplate of T4 and line parallel to superior endplate of L1*.
	Lumbar Cobb Angle (LC)	Angle subtended by line parallel to superior endplate of L1 and line parallel to inferior endplate of L3.

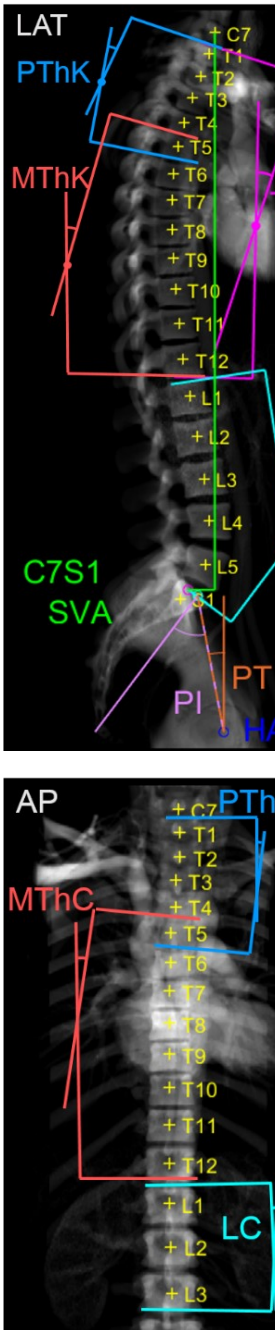

Figure 3.1. Definitions and illustration of various GSA metrics. Sagittal GSA is determined from LAT radiographs (top) and coronal GSA from PA or AP radiographs (bottom).

2.1 Manual Annotation

A reader study was performed to assess inter and intra-reader variability in manual annotation of vertebral endplate angles by five expert clinicians (two radiologists and three spinal neurosurgeons). Seven patient CT images were used from SpineWeb Dataset 14,[45] from which lateral (LAT) and anterior-posterior (AP) digitally reconstructed radiographs (DRRs) were computed spanning spinal vertebrae from C7 to S1. Readers annotated a single endplate digitally on each of 128 DRRs (49 LAT + 15 LAT repeats + 49 AP + 15 AP repeats).

The DRRs were generated with 0.6 mm pixel spacing and system geometry characterized by source-to-detector distance (SDD) = 120 cm and source-to-axis distance (SAD) = 60 cm using a prism projector (divergent fan-beams stacked in rows to form a full length DRR – thus simulating a line scanner or CT "scout" view). This projection model reduces distortion of endplates compared to a divergent cone-beam projection and represents an optimistic lower bound on variability in manual endplate delineation.

Reproducibility of manual annotation was quantified in terms of intraclass correlation coefficient (ICC), ranging from −1 to 1. Rubrics for ICC performance cited in Koo et al.[46] are 0.00–0.49 (poor), 0.50–0.74 (fair), 0.75–0.90 (good), and 0.90–1.00 (excellent). Inter-reader ICC was calculated as:

$$\mathrm{ICC}_{\mathrm{inter}} = \frac{\sigma_{\mathrm{WR}}^2}{\sigma_{\mathrm{BR}}^2 + \sigma_{\mathrm{WR}}^2} \tag{3.1}$$

where σ_{BR}^2 is the variance between five expert readers and σ_{WR}^2 is the mean variance within readers. Inter-reader ICC was calculated from 98 single endplate angle measurements (49 LAT + 49 AP). Intra-reader ICC was calculated for each reader as:

$$\mathrm{ICC}_{\mathrm{intra}} = \frac{\sigma_{\mathrm{WT}}^2}{\sigma_{\mathrm{BT}}^2 + (\mathrm{k}-1)\sigma_{\mathrm{WT}}^2} \tag{3.2}$$

where σ_{BT}^2 is the variance between repeated trials of measurements, σ_{WT}^2 is the variance between trials within a particular reader, and k is the number of repeated measurements. Intra-reader ICC was calculated from 60 single endplate angle measurements (15+15 LAT repeats + 15+15 AP repeats). Inter- reader root

mean square (RMS) was also computed as measures of the dispersion in endplate angle distribution within and across readers, and was used to compute the inter-reader 95% confidence interval (CI_{95}).

2.2 Automatic Method 1: Endplate Segmentation (EndSeg)

An automatic endplate segmentation method (denoted EndSeg) was developed to delineate endplate angles analogous to the manual method described in Section 2.1. With a CT image and vertebral labels (defined at the approximate centroid of the vertebral body) as input, the EndSeg algorithm identifies endplate angles as illustrated in Figure 3.2. A pseudocode description of operations is shown in Table 3.2. The input vertebral labels could be defined by any of the various automatic labeling methods described in Section 1.[18,20,37–40] In the work reported here, the vertebral point labels were defined manually by an expert radiologist at the approximate centroid of each vertebra. Sources of variation in the location of labels include intra-user variability (for manual labeling, as in this work) and image quality limitations (for manual or automatic labeling methods).

The EndSeg method starts with a segmentation of the vertebral body using continuous max-flow optimization,[47] which has demonstrated robust performance in a variety of applications, including spine imaging.[48] The max-flow objective function uses a maximum-a-posteriori estimate of voxel labels, a weighting function based-on the vertebral centroid seed point, and a smoothness preserving regularization term. Image erosion was applied to separate the vertebral body from the spinous process. Following erosion, principal component analysis of the larger component (vertebral body) was used to determine the direction of the principal axis of the vertebral body (roughly superior-inferior direction).

The EndSeg method then delineates the vertebral endplate starting with a seed point defined by the intersection of the principal axis with the vertebral body segmentation. A region-growing method expands about the seed point on the vertebral endplate, stopping near the edge of the endplate, where the gradient elevation direction (i.e., the angle of the surface out of the *x*-*y* plane) changes sharply. The region growing is constrained by the distance from the seed point and a gradient elevation threshold parameter. A voxel at

location x is added to the endplate segmentation if the gradient elevation at that location, $\nabla G(x)$, differs from the mean gradient elevation in a small patch around the seed point, $\overline{\nabla G}_e$ by less than the threshold, τ:

$$|\overline{\nabla G}_e - \nabla G(x)| < \tau \tag{3.3}$$

thus growing along the "flat" surface of the endplate. Region growing was constrained to the anterior portion of the vertebra (i.e., did not include the spinous or transverse processes) by a distance cutoff beyond the posterior surface of the vertebral body. The posterior constraint was defined using the principal component of the segmentation as a proxy for anterior-posterior axis. A sensitivity analysis was performed to investigate the dependence of EndSeg endplate angle estimation on the selection of the gradient threshold parameter, τ.

As illustrated in Figure 3.2, the EndSeg method then projects the locus of region-grown endplate voxels onto a DRR as a point-cloud distribution with forward projection as in Section 2.1. A linear fit to the point-cloud is then performed, and the slope is taken as a surrogate for the endplate angle. The resulting endplate angles are used as the basis to compute the various GSA metrics according to the definitions in Figure 3.1.

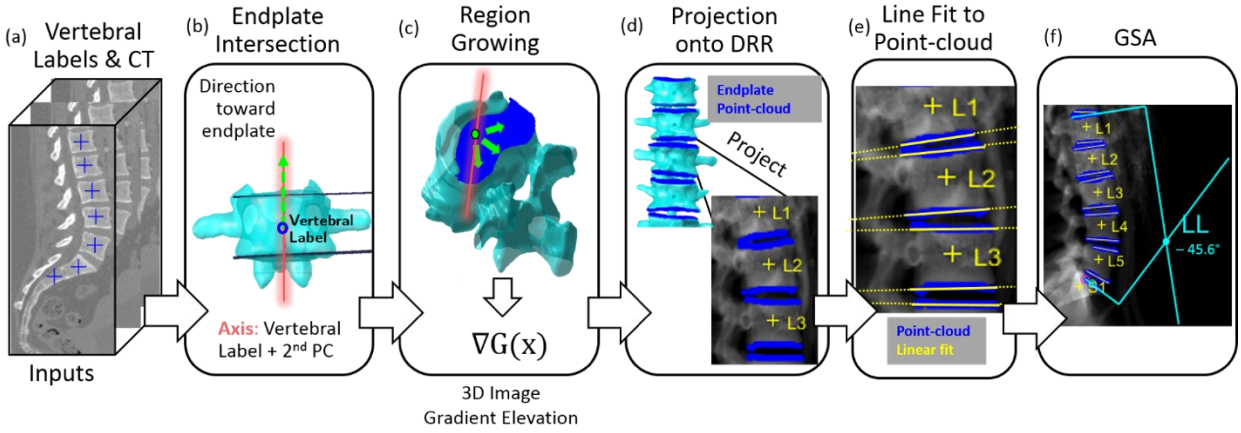

Figure 3.2. The EndSeg algorithm for calculation of GSA. (a) Input CT image with vertebral labels. (b) Computation of principal axis and intersection with endplate. (c) Region growing along the endplate. (d) Projection of the segmented endplate from 3D to the 2D LAT or AP radiographic plane. (e) Linear fit to the projected point cloud. (f) GSA metrics computed according to definitions in Figure 3.1.

Pseudocode	User Interaction
Get CT image, vertebral labels, and vertebra segmentation	
Get vertebral labels c_i	Manual
$c_i(x_i, y_i, z_i)$: 3D "centroid" position for vertebra i	or Auto [16-18]
for each vertebra $\boldsymbol{i}$:	
vertebra.seg = maxflow(CT, c_i)	Auto [30]
vertebra.seg = gaussian.filter(erosion(closing(vertebra.seg)))	Adjustable
Find endplate edge	
for each vertebra i:	
if not vertebra S1:	
principal.axes = PCA (vertebra.seg)	Auto [31]
principal.axis = max.z(principal.axes)	Auto
else:	
principal.axis = direction(c_i (S1) to c_{i+1} (L5))	Auto
endplate.edge = peak.gradient(principal axis)	Auto
Segment the endplate (region growing)	
specify τ: Angular threshold of region growing	Adjustable [32]
for each vertebra:	Auto
compute ∇G (local 3D image gradient)	
while $\nabla G < \tau$	
endplate.seg = endplate.seg + current voxel	Auto
endplate.seg = posterior.cutoff(endplate.seg)	Auto
Project endplate to DRR and compute endplate angle	
for each endplate.seg	
endplate.pointcloud = forward.projection(endplate.seg)	Auto
linear.fit = fit(endplate.pointcloud, 'linear')	Auto
endplate.angle = slope(linear.fit)	Auto

Table 3.2. Pseudocode for EndSeg. Functional blocks correspond to the Figure 3.2 flowchart. Calculations and parameters in the first column are denoted in the second column as manual, automatic, or adjustable.

2.3 Automatic Method 2: Spline-Fit Normals (SpNorm)

The Spline-Fit Normals (SpNorm) method computes spinal curvature without segmentation. Given vertebral labels as input (as in Section 2.2), SpNorm forward projects vertebral labels to a DRR and computes a 2D curvilinear fit as illustrated in Figure 3.3. A pseudocode description of operations is shown in Table 4. Various fit models were investigated, including a smoothing spline-fit and polynomial models (3rd, 5th, 6th, and 7th order). The smoothing-spline fit minimized the following objective for vertebral label coordinates (x_i, y_i) in the DRR:

$$s = \underset{s_j}{\operatorname{argmin}}\, p \sum_i \left(y_i - s_j(x_i)\right)^2 + (1-p)\int \left(\frac{d^2 s_j}{dx^2}\right)^2 dx \tag{3.4}$$

where s is the smoothing spline function, x_i is the distance along the cranial-caudal axis determined by orientation of the CT, y_i is distance perpendicular to the cranial-caudal axis determined by projection type (i.e., LAT, AP), and p is approximately $1/(1+\frac{h^3}{6})$, where h is the average data point spacing.

The SpNorm method then identifies rays normal to the spline as proxies for endplate angles. The normal ray intersection was nominally computed midway between vertebral levels, but locations varied somewhat depending on spinal region: half-way between the superior and inferior vertebral for levels C7-L5; and at a distance 0.65 of the way between labels from L5–S1 (slightly closer to S1 to more accurately capture the sharp curvature at the lumbosacral junction). The S2 level provided an additional inferior control point for curve fitting. Metrics of GSA were then computed as defined in Figure 3.1.

The sensitivity of SpNorm to variations in the location of the vertebral label (e.g., due to intra-user or image noise) was investigated by varying the label coordinate and analyzing the effect on GSA metric. Perturbations from the "true" (manually defined) locations were realized according to a Gaussian distribution with σ = 5.7 mm and a cutoff at Δx_j = 9 mm to confine within the vertebral body. A total of 100 perturbed locations were simulated for each level, and the resulting variation in GSA metric was analyzed.

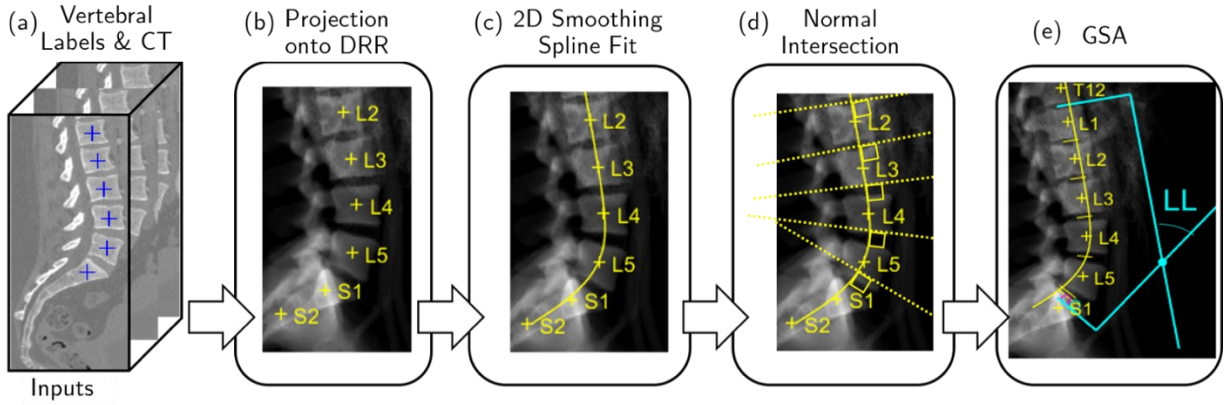

Figure 3.3. SpNorm algorithm for calculation of GSA. (a) Input CT image with vertebral labels. (b) Projection of vertebral labels from 3D to the 2D LAT or AP radiographic plane. (c) Spline fit to the projected labels. (d) Normal rays computed between vertebral labels, with slope taken as a proxy for endplate angle. (e) GSA metrics computed according to definitions in Figure 3.1.

Pseudocode	User Interaction
Get CT image and vertebral labels	
Get vertebral labels c_i	Manual
$c_i(x_i, y_i, z_i)$: 3D "centroid" position for vertebra i	or Auto [28]
Project labels to DRR	
for each vertebra $\boldsymbol{i}$	
$c_{R,i}(u_i, v_i) \leftarrow$ project label $c_i(x_i, y_i, z_i)$ to DRR	Auto
Fit smoothing spline	
spline.fit = fit($c_{R,i}$, 'smoothing_spline')	Auto
Compute spline normal	
specify $\boldsymbol{f_v}$: fraction of distance between vertebra to approximate	Adjustable
… endplate location	
specify $\boldsymbol{f_{S1}}$: fraction of distance between L5 and S1 to	Adjustable
… approximate S1 endplate location	
for each vertebra i	
if not vertebra L5	
spline.point = spline.fit($c_{R,i} + \boldsymbol{f_v} c_{R,i+1}$)	Auto
else	
spline.point = spline.fit($c_{R,L5} + \boldsymbol{f_{S1}} c_{R,S1}$)	Auto
spline.slope = derivative(spline.fit(spline.point))	Auto
Compute endplate angle	
endplate.angle = -1 / spline.slope	Auto

Table 3.3. Pseudocode for SpNorm. Functional blocks correspond to the Figure 3.3 flowchart. Calculations and parameters in the first column are denoted in the second column as manual, automatic, or adjustable.

2.4 Performance of Manual and Automatic Methods

The EndSeg and SpNorm methods were compared to manual annotation, which represents the conventional clinical means of GSA analysis and is therefore a reasonable basis of comparison (recognizing that it may or may not represent an actual "truth" definition). Endplate angles computed by EndSeg and SpNorm were compared to manual annotation using Passing-Bablok (PB) regression tests, a non-parametric test of similarity between measurements.[49] Deviations from the regression line were computed and compared to the CI_{95} of manual measurements. Such non-parametric measures provided insight for small sample sizes and spread about the median. Student's t-tests were used to assess differences between automatic and manual methods.

The performance of GSA measurement (from endplate angle estimates as described above) was similarly analyzed using PB regression tests to evaluate the similarity of manual and automatic methods. Metrics of

GSA included lumbar lordosis (LL), main thoracic kyphosis (MThK), proximal thoracic kyphosis (PThK), lumbar Cobb angle (LC), main thoracic Cobb angle (MThC), and proximal thoracic Cobb angle (PThC) – each defined in Figure 3.1. Adherence to the regression line within CI_{95} was tested, and parametric Student's t-tests were performed to compare the distributions in manually and automatically determined GSA metrics.

3 Results

3.1 Manual Annotation

Inter-reader and intra-reader ICC (ICC_{inter} and ICC_{intra}, respectively) are summarized in Table 3.4. The ICC_{inter} was similar to ICC_{intra} for both LAT and AP cases. Noting that ICC_{inter} is within two standard deviations of ICC_{intra}, it appears that visual interpretation of the endplate is no more consistent for a single reader than between different readers.

	ICC_{intra}						**ICC_{inter}**
	R1	**R2**	**R3**	**R4**	**R5**	**Average**	
AP	0.75 (0.39 - 0.91)	0.86 (0.73 - 0.93)	0.76 (0.60 - 0.86)	0.76 (0.62 - 0.85)	0.78 (0.67 - 0.86)	0.78 (± 0.05)	0.80 (0.69 - 0.87)
LAT	0.99 (0.96 - 1.00)	0.99 (0.98 - 1.00)	0.98 (0.95 - 0.99)	0.97 (0.94 - 0.98)	0.97 (0.95 - 0.98)	0.98 (± 0.01)	1.0 (0.99 - 1.00)

Table 3.4. Inter- and intra-reader agreement (ICC_{intra}, and ICC_{inter}) in manual definition of endplate angle. Parenthetical values denote the CI_{95} (or ± standard deviation for average ICC_{intra}).

According to Table 3.4, both ICC_{inter} and ICC_{intra} were greater in LAT views, suggesting that measures of sagittal curvature have higher reproducibility than coronal measures. This observation may be associated with better visualization (e.g., reduced overlap) of endplates for cases included in this study, which exhibited varying levels of normal or abnormal lordosis and kyphosis but did not include pathologically significant scoliosis.

However, the distributions in Figure 3.4 suggest that LAT views have slightly higher median and standard deviation in endplate angle definition than AP views. The apparent discrepancy between standard deviation and ICC is because the former solely describes inter/intra-reader variability. ICC additionally describes the

inherent range of variability within a single type of measurement (i.e., LAT views have a typical endplate angle range of −40 to 40°) as it is a ratio between the inherent variability in a particular type of measurement (σ_{WR} and σ_{WT}) and the sum of the inherent variability and inter- or intra-reader variability ($\boldsymbol{\sigma_{BR}}$ or $\boldsymbol{\sigma_{BT}}$, respectively). In the extreme case when inherent variability is much greater than the inter/intra-reader variability as with LAT views, then ICC approaches 1. From the standard deviations shown in Figure 3.4, inter-reader CI_{95} in endplate angle definition was 5.8°.

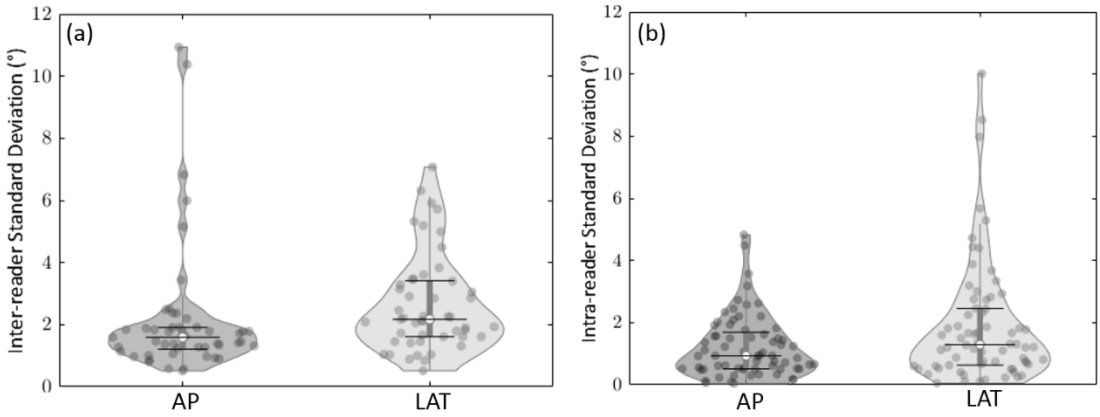

Figure 3.4. Standard deviation in (a) inter-reader and (b) intra-reader variability of endplate angle definition in AP and LAT views. The violin plots show individual sample points, an envelope fit to the sample distribution, the median (open circle with horizontal bar), and interquartile range (upper and lower horizontal range bars).

The inter-reader CI_{95} for sagittal GSA metrics was 8.2° for PThK, 6.0° for MThK, and 7.4° for LL, with a similar range for coronal GSA metrics: 11.0° for PThC, 8.2° for MThC, and 4.8° for LC. The mean inter-reader error for GSA metrics across all manual annotations was 7.6°. Because GSA metrics aggregate the error from two endplate angle measurements, they are subject to a higher rate of variability than the error for individual endplate measurements. The variability in GSA measurements are consistent with such propagation of error.

3.2 Automatic Method 1: Endplate Segmentation (EndSeg)

The sensitivity of EndSeg to selection of the gradient angle threshold parameter (τ) was analyzed to determine the operating range and a suitable, nominal parameter setting for the region growing operation. As shown in Figure 3.5, the error in endplate angle (calculated as the difference from manual annotation) was greater for low values of the threshold, $\tau < \sim 5°$. Above this level, endplate region growing appeared

stable, with error ~0.5°, which is within the range of errors associated with manual (inter- or intra-reader) delineation.

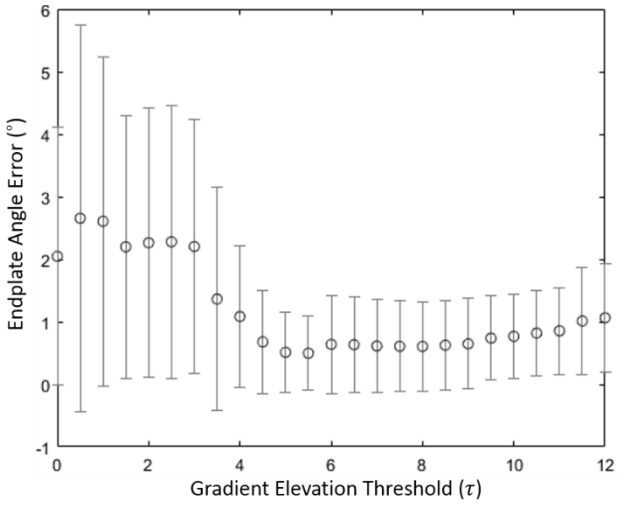

Figure 3.5. Sensitivity of endplate angle measurement in the EndSeg method to the gradient elevation threshold parameter, τ.

3.3 Automatic Method 2: Spline-Fit Normals (SpNorm)

Performance of various curvilinear fits is shown in Figure 3.6. As shown in Figure 3.6(a), the smoothing-spline model demonstrated the best adherence to manually defined endplate angles. The goodness of fit was quantified via the coefficient of determination (r^2) as shown in Figure 3.6(b), where the smoothing-spline model is seen to outperform other model types.

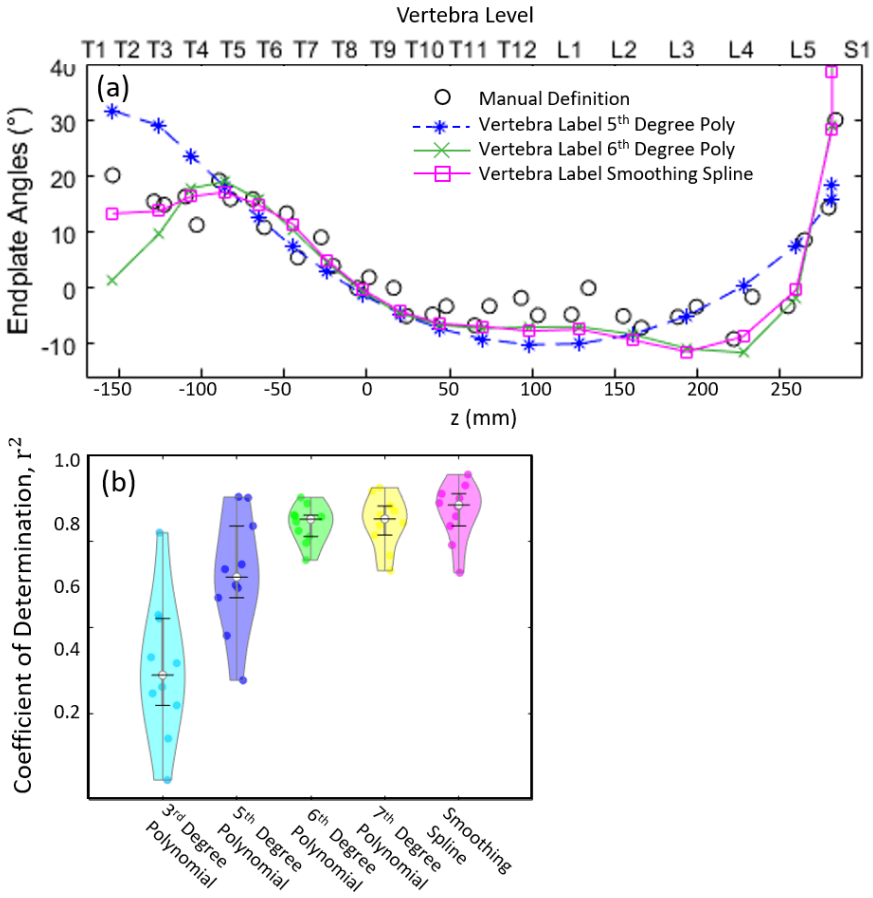

Figure 3.6. Selection of model fit for the SpNorm method. (a) Endplate angle evaluated for various model fits (solid curves) and by manual definition (open circle) shown as a function of position along the spine. (b) Coefficient of determination (r^2) between model-based and manual endplate angle definition for various models in SpNorm.

The sensitivity of SpNorm to variations in vertebral label location is shown in Figure 3.7(a), illustrating the impact of variation in the label coordinate on the spline fit. Relatively small deviations in the fit are observed with reasonable range of vertebral coordinate variations (e.g., within ±9 mm). Overall, 95% of the angle deviations are within the inter-reader CI_{95} as shown in Figure 3.7(b), with the exception of T1 and S1, for which the 95% bounds exceed the inter-reader CI_{95} (although the IQR was within the CI_{95}). The effect of label coordinate perturbations on GSA measurement is shown in Figure 3.7(c), with similar findings as for endplate angles (i.e., 95% of deviations within the inter-reader CI_{95}), with the exception of PThK and ThK.

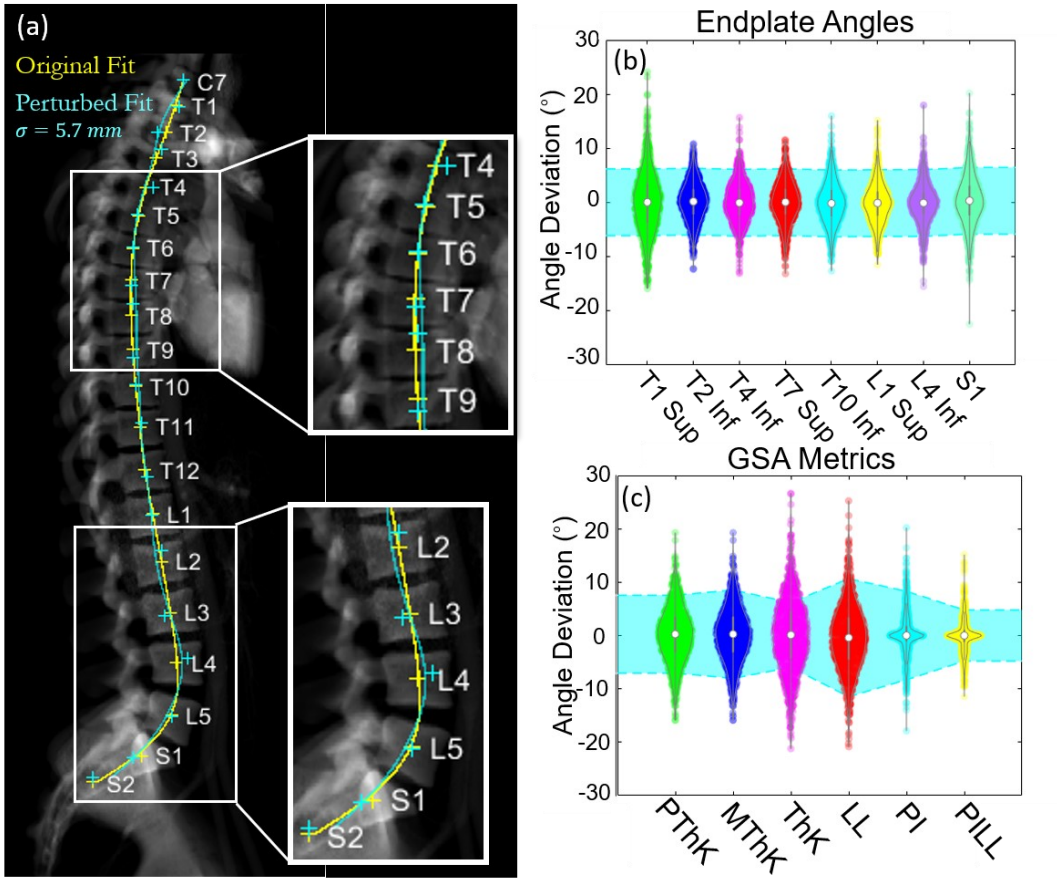

Figure 3.7. SpNorm sensitivity analysis. (a) LAT DRR with representative un-perturbed (yellow) and perturbed (cyan) SpNorm vertebral labels and fits. (b) Distribution in endplate angle estimation for SpNorm over the range of perturbation in vertebral label location. (c) Distribution in GSA metrics computed over the range of perturbed vertebral label location. The violin plots in (b) and (c) show the sample points, median (open circle), and interquartile range (vertical bar). The gray range in the background of (b) and (c) shows inter-reader CI_{95} for endplate angle and GSA metric, respectively.

3.4 Comparative Analysis of Manual and Automatic Methods

Figure 3.8 illustrates EndSeg and SpNorm GSA metric calculations. Most GSA metrics agreed within 3.0° (LL, PI, and PT) or ~10 mm (C7S1 SVA). The most notable difference was in ThK, with angle differences up to ~6°. The SpNorm estimate of ThK was unaffected by variations in endplate angle estimation and vertebral body shape variations as may affect EndSeg.

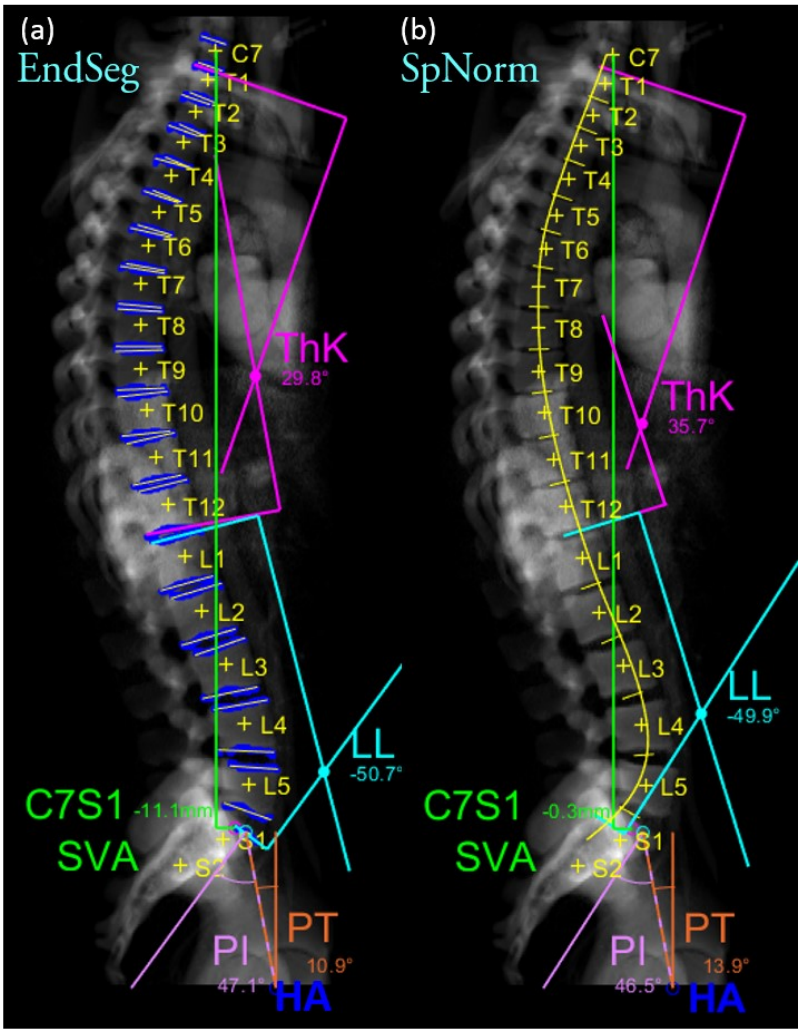

Figure 3.8. Example LAT DRR with GSA metrics as measured by (a) EndSeg and (b) SpNorm.

3.4.1 *Endplate Angles*

Endplate angle estimation for manual, EndSeg, and SpNorm methods suggests similar underlying distributions as seen in Figure 3.9. The alternative hypothesis (H_A) that the methods sample from distinct distributions ($p < 0.05$) was rejected for PB regression tests (manual to EndSeg, manual to SpNorm). Deviations in endplate angles from the regression line show that 93.8% of the endplate angle estimates for both EndSeg and SpNorm are within the inter-reader CI_{95}. With the exception of three outliers (> 2 standard deviations) for each method, endplate angle estimates suggested strong correspondence with manual definition. Comparison of example endplate angle measurements in Figure 3.9(c-d) showed comparable method performance, with no visible outliers.

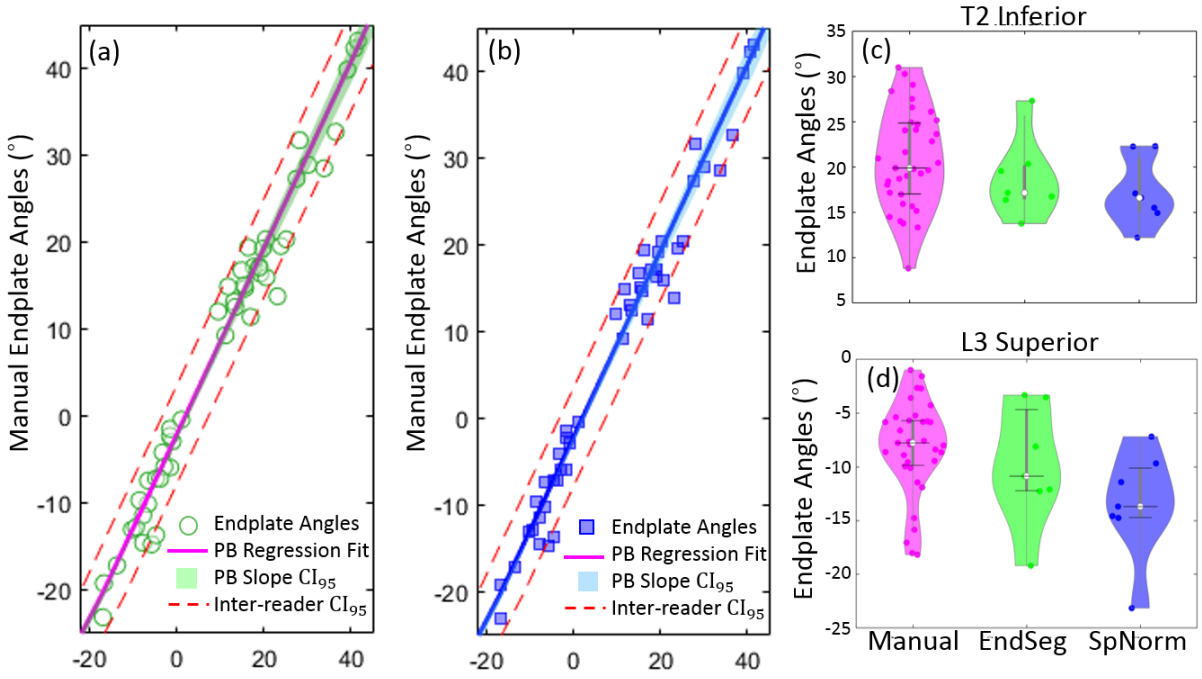

Figure 3.9. Comparison of automatic and manual endplate angle estimation. (a) Endplate angle PB regression between Manual and EndSeg, and (b) between Manual and SpNorm. The narrow transparent range marked about the fit shows the CI_{95} in PB regression slope, and the dashed lines mark the CI_{95} for inter-reader variability. Example endplate angle measurements for Manual, EndSeg, and SpNorm are shown in (c) for the T2 inferior endplate and in (d) for the L3 superior endplate. (Other endplates showed similar trends.) The violin plots show sample points, envelope of the distribution, median (open circle) and IQR.

Table 3.5 summarizes the average endplate angle difference between EndSeg and SpNorm compared to manual definition. EndSeg exhibited stricter adherence to manual definition than SpNorm in lower thoracic/lumbar regions, potentially due to the interplay between endplate plateaus and the curvature line along the spine, whereas endplate plateaus are more normal to the direction of curvature in the upper spine. Although EndSeg was observed to have an average absolute difference in endplate angle measurement less than SpNorm, both methods were within inter-reader CI_{95}. Student's t-tests failed to detect significant differences between manual and automatic methods.

	T2	T3	T4	T8	L1	L3	S1	Avg. Diff. (abs. val.)	T-test (p-value)
EndSeg	-3.6°	-4.0°	-2.2°	0.1°	0.4°	1.9°	0.8°	1.9°	p = 0.33
SpNorm	-2.1°	-1.6°	-0.4°	1.4°	2.6°	5.6°	-5.3°	2.7°	p = 0.83

Table 3.5. Difference in average endplate angle for EndSeg and SpNorm compared to manual definition for various vertebral levels. Paired T-tests reject the hypothesis that automatic and manual measurements are different for both EndSeg and SpNorm.

3.4.2 GSA Metric Computation

Pairwise Student's t-tests and PB regression tests (Figure 3.10) between manual and automatic GSA measurement methods (EndSeg, SpNorm) rejected the alternative hypothesis (H_A), thus failing to detect statistically significant differences between methods. All deviations of GSA metrics from the regression line are within inter-reader CI_{95}. Most notable differences in GSA metrics were observed in the sagittal upper thoracic measures (PThK in Figure 3.10(a) and MThK in Figure 3.10(b)), but strong correspondence was seen between automatic and manual GSA metric estimates. Violin plots of GSA metric examples in Figure 3.10 suggest similar GSA metric estimates for the three methods, although significant outliers (> 2 standard deviation) were observed in manual definition for MThC and PThC, which are not present in either automatic method.

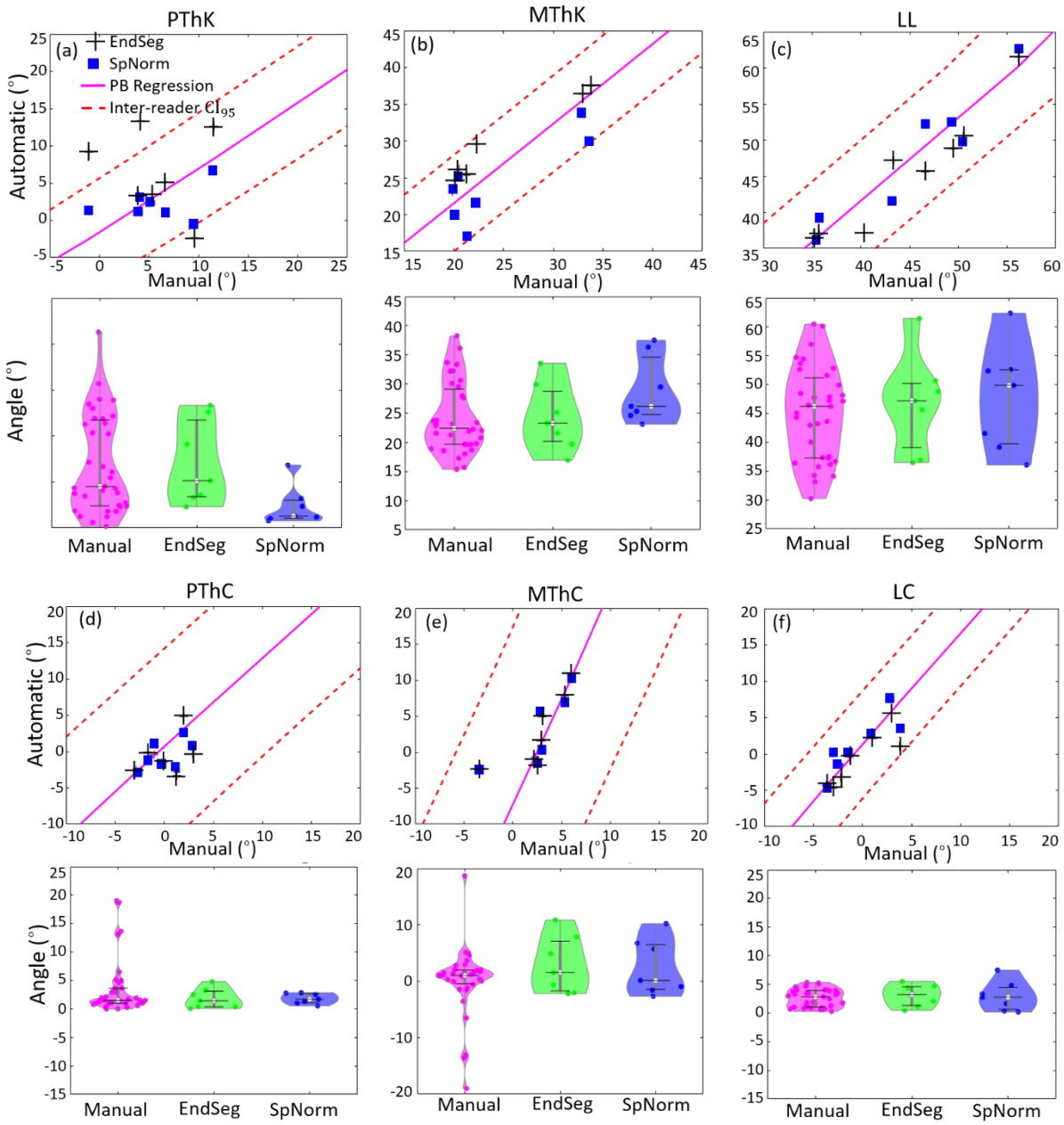

Figure 3.10. GSA metric comparison for automatic and manual methods for (a–c) sagittal and (d–f) coronal metrics of spinal alignment. In each case, the PB regression shows correspondence between automatic and manual methods as in Figure 3.9. Violin plots show sample points, distribution, median, and IQR in GSA metrics for each method.

Figure 3.11 shows PB regression between EndSeg and SpNorm for GSA metrics that require hip axis annotation. All tests rejected the alternative hypothesis (H_A) and failed to indicate significant differences between GSA metric estimates. Correspondence between the two methods suggests extension of EndSeg

and SpNorm to pelvic measures of alignment, although manual definitions of hip axis were not collected in the current work.

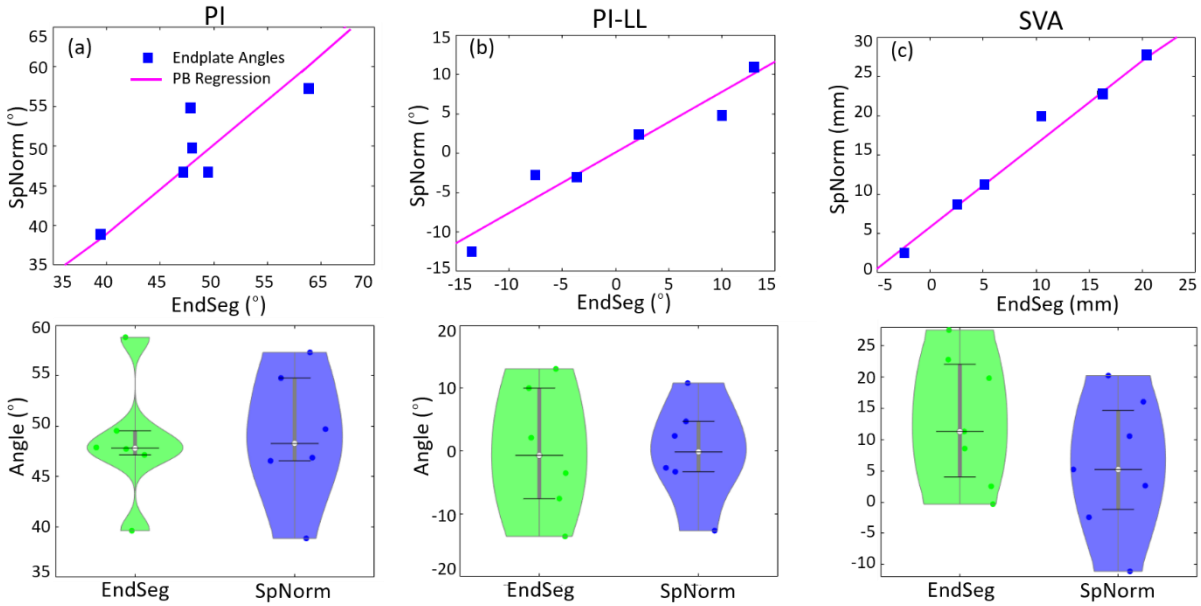

Figure 3.11. Comparison of GSA metric estimated by EndSeg and SpNorm for (a) PI, (b) PI-LL, and (c) SVA. The PB regression plots show correspondence in GSA metrics between the two methods, and violin plots show the sample points, distribution, median, and IQR in GSA estimate for each case.

4 Discussion and Conclusions

Two methods for automatic analysis of GSA from spinal CT were presented in this chapter – the first (EndSeg) based on endplate angle definition analogous to conventional endplate visualization and the second (SpNorm) based on angles normal to a robust spline fit of vertebral body labels. Both take vertebral body annotations defined in CT as basic input, and each projects to a 2D radiographic plane to analyze GSA in terms common to conventional analysis (as defined in Figure 3.1). EndSeg and SpNorm measurements demonstrated that 93.8% of endplate angles and all GSA metrics were within the inter-reader CI_{95} of manual measurements. There was no significant difference between GSA metrics determined by the manual and automatic methods.

Such automatic methods could help to avoid the high level of variability observed in manual definition (Table 2 and Refs. [14,32,33,35]) and could streamline efficiency compared to fairly time-consuming manual

analysis. Such capability could benefit more widespread analysis of GSA (e.g., retrospectively and/or intraoperatively) in large datasets and improve understanding of correlation between GSA and surgical outcomes. Such methods could be applied to intraoperative CT at the end of a case for validation of the surgical construct, with automatic analysis of the change in GSA imparted by surgery. In contrast, manual annotation disrupts surgical workflow and is not commonly practiced in intraoperative evaluation. With the advent of image analytic tools for correlative analyses of patient outcomes and risk factors, metrics computed by such automated method could help to glean important correlates and drive improvement in spinal surgery outcomes.

An important consideration for future work is the extension to more patient cases. This work describes application of both automatic algorithms to seven SpineWeb patient cases with similar scan protocols and patient pathologies, subsequently limiting the scope of the results. For future work, broader variation in CT scan protocol or patient pathology would permit validation of the algorithms for potentially broader application. The increased statistical power that accompanies an increase in the number of cases could elucidate important differences / similarities between methods that are not evident from this work. While SpNorm and EndSeg demonstrate similar performance within the current dataset, SpNorm is potentially more robust to vertebral shape abnormalities, owing to its shape-agnostic description of the curvature of the spine. Investigation of GSA analysis in pathologic spinal cases (e.g., ankylosing spondylitis or butterfly vertebra) could demonstrate further utility of SpNorm compared to endplate-based methods, in which pathologic vertebral shapes can alter representations of global spinal curvature.

Although EndSeg operates within a vertebral segmentation and both EndSeg and SpNorm require vertebral labels as input, existing methods for such inputs are numerous.[18,20,37–40] In the future, EndSeg will be investigated in unsegmented CT (i.e., operating on image gradients rather than segmentation boundaries), thus eliminating the need for segmentation prior to GSA measurements. Variations in CT scan protocols are particularly relevant, since poor image quality (e.g., large slice thickness) can affect the accuracy of vertebral segmentation.

Inter- and intra-reader variability reported in Section 3.1 quantify the reliability and reproducibility in DRRs alone. The measurements represent conservative lower bounds, since DRRs lack pertinent forms of image noise and artifacts that can challenge anatomical landmark identification in true radiographs. Thus, the respective measures are likely lower bounds on manual variability, and do not quantify all sources of variability present in radiographic measures. Comparison of automatically derived measures to manual ranges of variability serves as a preliminary baseline for validation in the current work.

The conventional approach to GSA analysis involves manual annotation of radiographs directly, whereas the current work applied EndSeg and SpNorm to radiographic GSA analysis from a CT image. Preoperative assessment of the patient commonly includes a CT image, while intraoperative and postoperative assessment is frequently limited to radiographs. Additional consideration of the effect of patient positioning on spinal curvature is important, recognizing that diagnostic radiographs are typically acquired under load-bearing conditions, whereas CT is typically acquired in supine or prone positioning. Although findings have demonstrated a correlation between metrics of GSA between standing, supine, and prone positioning, direct comparison may be difficult. Direct analysis of GSA from radiographs in future work could more readily allow computation of perioperative changes in GSA.

Chapter 4: Discussions and Conclusion

This thesis presented image analysis techniques to automatically localize surgical instrumentation and changes in spinal curvature in intraoperative radiographic imaging. Chapter 2 covered an algorithm to automatically detect and localize pedicle screws in intraoperative radiographs using a deep learning framework. The network achieved fairly high precision (~92.1%) and recall (~88.8%) in screw detection for neural networks trained on multiple-patient training and testing sets. Such an algorithm could improve the capture range and runtime of registration methods such as KC-Reg by providing a reliable initialization of screw locations. Chapter 3 reported the development of two algorithms to automatically analyze metrics of GSA in CT/CBCT. The two methods (EndSeg and SpNorm) demonstrated accurate assessment of GSA, with 93.8% of measurements within the CI_{95} of manual measurements. Both chapters present automatic/semi-automatic methods with sufficient accuracy to provide meaningful quantitative intraoperative QA. Work is underway to use such methods as a basis for data-intensive, retrospective analysis of large image datasets to extract image features that may correlate with surgical outcome.[15]

Recent advances in spine surgery have begun to integrate the use of machine learning to perform automatic detection and labeling of spine imaging. Localizing and labeling spine structures has achieved excellent performance comparable to expert annotation. Such methodology integrated with the electronic medical record (EMR) and Picture and Archiving Communication Systems (PACS)[50] could improve clinical decision support in diagnosis of spinal pathology and correlating image information with pain or functional outcomes.[51] Extending these methods to detection and characterization of medical implants, such as the approach covered in Chapter 2, can similarly provide quantitative input to determining the proper course of treatment or rehabilitation.

Additional forms of image processing have arisen with the increased prevalence of machine learning. Automatic anatomical segmentation has been a rapidly advancing field, including automatic segmentation of the spine.[18,37,52–59] Differences in vertebra shape due to anatomical deformities and degenerative pathologies challenge such algorithms and necessitate manual confirmation by the clinician. As covered in

Chapter 3, segmentation of vertebrae could provide useful input to clinical decision support based on vertebral shape and intervertebral spacing.[50]

Other important image-analytic methods emerging in recent research include automatic planning of pedicle screw trajectories using an atlas-based registration of anatomy and reference trajectories[60] and classification of scoliotic curve severity and location to determine the risk of curve progression.[61] Taken together, these tools could offer substantial improvements in patient care pathways, help to reduce medical error, increase the precision in the administration of spine surgery, and ultimately improve patient outcomes.

As described in Chapter 2, automatic detection of surgical spine surgery instrumentation helps to better integrate the KC-Reg framework into intraoperative workflow, obviating the need for manual input, and increasing the capture range via robust initialization. This method was developed using a network trained on radiographic images across four patients scanned on the Medtronic O-arm with simulated instrumentation using Medtronic Solera spine screw models. Generalization of network detection may be affected by system geometry, the appearance of the spine screws, and variations in patient anatomy. Using a larger set of training data would improve the generalizability and is the subject of future work. The method could be further generalized to detect different types of instrumentation, as well as learning to differentiate them – e.g., classification of surgical implant, surgical retractor, rod, or other medical devices can provide further clinical decision support. These techniques for detection of surgical implants have utility beyond spine surgery and could provide a basis for more generally applicable instrumentation classifiers for radiographing imaging.

Chapter 3 presented two methods for automatic analysis of GSA: one using seed points in CT (SpNorm); and the other using vertebra segmentations in CT (EndSeg). Reliable and accurate algorithms exist to automatically label vertebrae,[37–40,50] providing an accurate starting point for SpNorm. Although the current algorithm operates in CT, the approach is also compatible with intraoperative CT and CBCT, which supports its application in surgery. Extending the method from CT to radiography could be accomplished by a variety of 3D-2D registration techniques, recognizing challenges associated with deformation of the

spine between preoperative CT and intraoperative radiographs. Multi-stage 3D-2D deformable registration as proposed by Ketcha et al,[62] could account for spinal deformation while maintaining the rigid structure of the vertebrae. Application of this method could provide access to intraoperative QA without the need for intraoperative CT. Additional questions arise in comparing GSA assessed in images acquired under load-bearing conditions – as in a standing radiograph – to non-load-bearing conditions – as in a supine / prone CT, although correlation between the two has been demonstrated.[63]

The methods developed in this thesis advanced two key examples of image-analytic techniques for improving the accuracy, precision, and quality assurance of spine surgery. For the screw detection algorithm, as mentioned above, future work should include larger training datasets and extend the detection methods to other types of instrumentation. Similarly, extension of the automatic GSA analysis algorithm (particularly SpNorm) to automatically compute other potentially important image features – e.g., local curvature and intervertebral space – is a topic for future development. In both contexts, the extension of the methods from current stages to more generalizable forms with efficient, integrated algorithms is essential to eventual clinical translation.

Quantitative features derived from image data appear to be an important factor in improving the performance of clinical outcomes prediction, as demonstrated in De Silva et al.[15] Future integration of image analytics within healthcare interfaces to electronic medical records will be vital to realizing such potential in the shared decision support process between surgeons and patients. Considering the high prevalence of spine surgery and the fairly broad variability in surgical outcomes within existing techniques, even small gains in clinical decision support could have major benefit to many patients, and the promise of improved quality of care presents an important driving force for progress.

Bibliography

[1] Raciborski, F., Gasik, R. and Klak, A., "Disorders of the spine. A major health and social problem," Reumatologia **54**(4), 196–200 (2016).

[2] Daniell, J. R. and Osit, O. L., "Failed Back Surgery Syndrome: A Review Article," Asian Spine J **12**(2), 372–379 (2018).

[3] Martin, B., Mirz, S., Spin, N., Spiker, W., Lawrence, B. and Brodke, D., "Trends in Lumbar Fusion Procedure Rates and Associated Hospital Costs for Degenerative Spinal Diseases in the United States, 2004 to 2015.," Spine (Phila. Pa. 1976). **44**(5), 369–376 (2019).

[4] Mody, M., Nourbakhsh, A., Stahl, D., Gibbs, M., Alfawareh, M. and Garges, K., "The prevalence of wrong level surgery among spine surgeons.," Spine (Phila. Pa. 1976). **33**(2), 194–198 (2008).

[5] Rahmathulla, G., Nottmeier, E. W., Pirris, S. M., Gordon Deen, H. and Pichelmann, M. A., "Intraoperative image-guided spinal navigation: Technical pitfalls and their avoidance," Neurosurg. Focus **36**(3), 1–14 (2014).

[6] Mezger, U., Jendrewski, C. and Bartels, M., "Navigation in surgery," Langenbecks Arch Surg. **398**(4), 501–514 (2013).

[7] Kim, T. T., Johnson, J. P., Pashman, R. and Drazin, D., "Minimally Invasive Spinal Surgery with Intraoperative Image-Guided Navigations," Biomed Res. Int. **2016**(Special Issue) (2016).

[8] Yeramaneni, S., Robinson, C. and Hostin, R., "Impact of spine surgery complications on costs associated with management of adult spinal deformity," Curr. Rev. Musculoskelet. Med. **9**(3), 327–332 (2016).

[9] Uneri, A., De Silva, T., Goerres, J., Jacobson, M. W., Ketcha, M. D., Reaungamorhrat, S., Kleinszig, G., Vogt, S., Khanna, A. J., Osgood, G. M., Wolinsky, J.-P. and H, S. J., "Intraoperative evaluation of device placement in spine surgery using known-component 3D–2D image registration," Phys. Med. Biol. **62**(8), 3330–3351 (2017).

[10] Schwab, F., Patel, A. and Ungar, B., "Adult spinal deformity postoperative standing imbalance: how much can you tolerate? An overview of key parameters in assessing alignment and planning corrective surgery," Spine (Phila. Pa. 1976). **35**, 2224–2231 (2010).

[11] Glassman, S., Bridwell, K. and Dimar, J., "The impact of positive sagittal balance in adult spinal deformity.," Spine (Phila. Pa. 1976). **30**, 2024–2029 (2005).

[12] Diebo, B., Oren, J., Challier, V., Lafage, R., Ferrero, E., Liu, S., Vira, S., Spiegel, M., Harris, B., Liabaud, B., Henry, J., Errico, T., Schwab, F. and Lafage, V., "Global sagittal axis: a step toward full-body assessment of sagittal plane deformity in the human body," J Neurosurg Spine **25**(4), 494–499 (2016).

[13] Leveque, J.-C. A., Segebarth, B., Schroerlucke, S., Khanna, N., Pollina, J. J., Youssef, J., Tohmeh, A. and Uribe, J., "A multicenter radiographic evaluation of the rates of preoperative and postoperative malalignment in degenerative spinal fusions," Spine (Phila. Pa. 1976). **43**(13), E782–E789 (2018).

[14] Dang, N. R., Moreau, M. J., Hill, D. L., Mahood, J. K. and Raso, J., "Intra-observer reproducibility and interobserver reliability of the radiographic parameters in the spinal deformity study group's AIS radiographic measurement manual.," Spine. **30**(9), 1064–1069 (2005).

[15] De Silva, T. S., Vedula, S. S., Perdomo-Pantoja, A., Vijayan, R. C., Doerr, S. A., Uneri, A., Han, R., Ketcha, M. D., Skolasky, R. L., Witham, T., Theodore, N. and Siewerdsen, J. H., "SpineCloud: image analytics for predictive modeling of spine surgery outcomes," J. Med. Imaging **7**(3), 031502 (2020).

[16] Bortfeld, T., "Optimized Planning Using Physical Objectives and Constraints," Semin. Radiat. Oncol. **9**(1), 20–34 (1999).

[17] Lehman, R. A., Lenke, L. G., Keeler, K. A., Kim, Y. J. and Cheh, G., "Computed tomography evaluation of pedicle screws placed in the pediatric deformed spine over an 8-year period," Spine (Phila. Pa. 1976). **32**(24), 2679–2684 (2007).

[18] Levine, M., De Silva, T., Ketcha, M., Vijayan, R., Doerr, S., Uneri, A., Vedula, S., Siewerdsen, J. and Theodore, N., "Automatic vertebrae localization in spine CT: a deep-learning approach for image guidance and surgical data science," SPIE Med. Imaging **10951**, 109510S (2019).

[19] Suzani, A., Seitel, A., Liu, Y., Fels, S. and Roholing, R. N., "Fast automatic vertebrae detection and localization in pathological CT scans – a deep learning approach," MICCAI, 678–686, Springer International Publishing (2015).

[20] Yang, D., Xiong, T., Xu, D., Huang, Q., Liu, D., Zhou, S. K., Xu, Z., Park, J., Chen, M., Tran, T. D., Chin, S. P. and Metaxas, D., "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network to message passing and sparsity regularization," Inf. Process. Med. Imaging **10265**, 633–644, Springer International Publishing (2017).

[21] Kügler, D., Jastrzebski, M. A. and Mukhopadhyay, A., "Instrument Pose Estimation Using Registration for Otobasis Surgery," Biomed. Image Regist. **10883**, 105–114 (2018).

[22] Uneri, A., Zhang, X., Yi, T., Stayman, J. W., Helm, P. A., Osgood, G. M., Theodore, N. and Siewerdsen, J. H., "Known-component metal artifact reduction (KC-MAR) for cone-beam CT," Phys. Med. Biol. **64**(16), 165021 (2019).

[23] Zhang, X., Uneri, A., Webster Stayman, J., Zygourakis, C. C., Lo, S. L., Theodore, N. and Siewerdsen, J. H., "Known-component 3D image reconstruction for improved intraoperative imaging in spine surgery: A clinical pilot study," Med. Phys. **46**(8), 3483–3495 (2019).

[24] Zhao, Z., Zheng, P., Xu, S. and Wu, X., "Object Detection With Deep Learning: A Review," IEEE Trans. Neural Networks Learn. Syst. **30**(11), 3212–3232 (2019).

[25] Ren, S., He, K., Girshick, R. and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," Adv. Neural Inf. Process. Syst., 91–99 (2015).

[26] Wang, A. S., Stayman, J. W., Otake, Y., Vogt, S., Kleinszig, G., Khanna, A. J., Gallia, G. L. and Siewerdsen, J. H., "Low-dose preview for patient-specific, task-specific technique selection in cone-beam CT," Med. Phys. **41**(7), 071915 (2014).

[27] Nakai, S., Yoshizawa, H. and Kobayashi, S., "Long-term follow-up study of posterior lumbar interbody fusion," J. Spinal Disord. **12**(4), 293–299 (1999).

[28] Fon, G. T., Pitt, M. J. and Thies, A. C., "Thoracic kyphosis: range in normal subjects," Am. J. Roentgenol. **134**(5), 979–983 (1980).

[29] Tempel, Z., Gandhoke, G., Bolinger, B., Khattar, N., Parry, P., Chnag, Y., Okonkwo, D. and Kanter, A., "The influence of pelvic incidence and lumbar lordosis mismatch on development of symptomatic adjacent level disease following single-level transforaminal lumbar interbody fusion," Neurosurgeryj **80**(6), 880–886 (2017).

[30] Kyrola, K., Salme, J., Tuija, J., Tero, I., Eero, K. and Arja, H., "Intra- and interrater reliability of sagittal spinopelvic parameters on full-spine radiographs in adults with symptomatic spinal disorders.," J Neurospine. **15**(2), 175–181 (2018).

[31] Carman, D., Browne, R. and JG, B., "Measurement of scoliosis and kyphosis radiographs. Intraobserver and interobserver variation.," J Bone Jt. Surg Am. **72**(3), 328–333 (1990).

[32] Vaynrub, M., Hirsch, B. P., Tishelman, J., Dennis, V.-M., J, Bu. A., Errico, T. J. and Protopsaltis, T. S., "Validation of prone intraoperative measurements of global spinal alignment.," J Neurosurg Spine. **29**(2), 187–192 (2018).

[33] Dimar, J., Carreon, L., Labelle, H., Djurasovic, M., Weidenbaum, M., Brown, C. and Roussouly, P., "Intra- and inter-observer reliability of determining radiographic sagittal parameters of the spine and pelvis using a manual and a computer-assisted method," Eur Spine J **17**(10), 1373–1379 (2008).

[34] Orht-Nissen, S., Cheung, J. P. Y., Hallager, D. W., Cehrchen, M., Kwan, K., Dahl, B., Cheung, K. M. and Samartzis, D., "Reproducibility of thoracic kyphosis measurements in patients with adolescent idiopathic scoliosis," Scolisosi Spinal Disord **12**(4), 1–8 (2017).

[35] Yamada, K., Aota, Y. and Higashi, T., "Accuracies in measuring spinopelvic parameters in full-spine lateral standing radiograph," Spine. **40**(11), E640-6 (2015).

[36] Wu, W., Liang, J., Du, Y., Tan, X., Xiang, X., Wang, W., Ru, N. and Le, J., "Reliability and reproducibility analysis of the Cobb angle and assessing sagittal plane by computer-assisted and manual measurement tools," BMC Musculoskelet. Disord. **15**(33) (2014).

[37] Klinder, T., Ostermann, J., Ehm, M., Franz, A., Knesr, R. and Lorenz, C., "Automated model-based vertebra detection, identification, and segmentation in CT images," Med Image Anal **13**(3), 471–482 (2009).

[38] Naegel, B., "Using mathematical morphology for the anatomical labeling of vertebrae from 3D CT-scan images," Comput Med Imaging Graph **31**(3), 141–156 (2007).

[39] Glocker, B., Feulner, J., Criminist, A., Haynor, D. and Konukoglu, E., "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans," Int. Conf. Med. Image Comput. Comput. **7512**, 590–598 (2012).

[40] Qadri, S. F., Ai, D., Guoyu, H., Ahmad, M., Huang, Y., Wang, Y. and Yang, J., "Automatic deep feature learning via patch-based deep belief network for vertebrae segmentation in CT images," MDPI **9**(1), 69 (2018).

[41] Ailon, T., Scheer, J., Lafage, V., Schwab, F., Klineberg, E., Sciubbia, D., Protopsaltis, T., Zebala, L., Hostin, R., Obeid, I., Koski, T., Kelly, M., Bess, S., Shaffrey, C., Smith, J., Ames, C. and Group, I. S. S., "Adult spinal deformity surgeons are unable to accurately predict postoperative spinal alignment using clinical judgment alone," Spine Deform **4**(4), 323–329 (2016).

[42] Bourgeois, A. C., Faulkner, A. R., Pasciak, A. S. and Bradley, Y. C., "The evolution of image-guided lumbosacral spine surgery," Ann Transl Med **3**(5), 69 (2015).

[43] Holly, L. and Foley, K., "Image guidance in spine surgery," Orthop Clin North Am **38**(3), 451–461 (2007).

[44] Tjardes, T., Shafizadeh, S., Rixen, D., Paffrath, T., Bouillon, B., Steinhausen, E. and Baethis, H., "Image-guided spine surgery: state of the art and future directions," Eur Spine J **19**(1), 25–45 (2010).

[45] Muñoz, H. E., Yao, J., Burns, J. E. and Summers, R. M., "Detection of vertebral degenerative disc disease based on cortical shell unwrapping," SPIE Med. Imaging, 2013 **8670**, 86700C, Springer Berlin Heidelberg, Berlin, Heidelberg (2013).

[46] Koo, T. and Li, M., "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," J Chiropr Med **15**(2), 155–163 (2016).

[47] Iquebal, A. S. and Bukkapatnam, S. T. S., "Unsupervised image segmentation via maximum a posteriori estimation of continuous max-flow," CoRR **abs/1811.00220** (2018).

[48] Pezold, S., Fundana, K., Amann, M., Andelova, M., Pfister, A., Till, S. and Cattin, P. C., "Automatic Segmentation of the Spinal Cord Using Continuous Max Flow with Cross-sectional Similarity Prior and Tubularity Features," [Recent Advances in Computational Methods and Clinical Applications for Spine Imaging], J. Yao, B. Glocker, T. Klinder, and S. Li, Eds., Springer International Publishing, Cham, 107–118 (2015).

[49] Passing, H. and Bablok, W., "A new biometrical procedure for testing the equality of measurements from two different analytical methods," J Clin CHem Clin Biochem **21**(11), 709–720 (1983).

[50] Galbusera, F., Casaroli, G. and Bassani, T., "Artificial intelligence and machine learning in spine research," JOR Spine **2**(1), e1044 (2019).

[51] Bounds, D., Lloyd, P., Mathew, B. and Waddell, G., "A multilayer perceptron network for the diagnosis of low back pain.," Proc. IEEE Int. Conf. Neural Networks **2**, 481–489 (1988).

[52] Law, M., Tay, K., Leung, A., J Garvin, G. and Li, S., "Intervertebral disc segmentaiton in MR images using ansiotropic oriented flux," Med Image Anal **17**(1), 43–61 (2013).

[53] Michopoulou, S., Costaridou, L., Panagiotopouos, E., Speller, R., Panayiotakis, G. and Todd-Pokropek, A., "Atlas-base segmentation of degenerated lumbar intervetebral discs from MR images of the spine.," IEEE Trans Biomed Eng **56**(9), 2225–2231 (2009).

[54] Neubert, A., Fripp, J., Engstrom, C., Schwarz, R., Lauer, L., Salvado, O. and Crozier, S., "Automated detection, 3D segmentation and analysis of high resolution spine MR images using statistical shape models," Phys Med Biol **57**(24), 8357–8376 (2012).

[55] Korez, R., Ibragimov, B., Likar, B., Pernus, F. and Vrtovec, T., "Deformable model-based segmentation of intervertebral discs from MR spine images by using the SSC descirptor," Comput. Methods Clin. Appl. Spine Imaging **9402**, 117–124 (2016).

[56] Ayed, I., Punithakumar, K., Garvin, G., Romano, W. and Li, S., "Graph cuts with invariant object-interaction priors: application to intervertebral disc segmentation," Inf Process Med Imaging **6801**, 221–232, Springer Berlin Heidelberg (2011).

[57] Carballido-Gamio, J., Belongie, S. and Majumdar, S., "Normalized cuts in 3-D for spinal MRI segmentation," IEEE Trans Med Imaging **23**(1), 36–44 (2004).

[58] Egger, J., Kapur, T., Dukatz, T., Kolodziej, M., Zukic, D., Freisleben, B. and Nimsky, C., "Square-cut: a segmentation algorithm on the basis of a rectangle shape.," PLoS One **7**(2), e31064 (2012).

[59] Huang, S., Chu, Y., Lai, S. and Novak, C., "Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI," IEEE Trans Med Imaging2 **28**(10), 1595–1605 (2009).

[60] Vijayan, R., De SIlva, T., Han, R., Zhang, X., Uneri, A., Doerr, S., Kethca, M., Perdomo-Pantoja, A., Theodore, N. and Siewerdsen, J., "Automatic pedicle screw planning using atlas-based registration of anatomy and reference trajectories.," Phys Med Biol **64**(16), 165020 (2019).

[61] Komeili, A., Westover, L., Parent, E., El-Rich, M. and Adeeb, S., "Monitoring for idiopathic scoliosis curve progression using surface topography asymmetry analysis of the torso in adolescents," Spine J **15**(4), 743–751 (2015).

[62] Ketcha, M., De Silva, T., Uneri, A., Jacobson, M., Goerres, J., Kleinszig, G., Vogt, S., Wolinsky, J. and Siewerdsen, J., "Multi-stage 3D-2D registration for correction of anatomical deformation in image-guided spine surgery," Phys. Med. Biol. **62**(11), 4606–4622 (2017).

[63] Brink, R. C., Colo, D., Schlosser, T. P., Vincken, K. L., van Stralen, M., Hui, S. C., Shi, L., Chu, W. C., Cheng, J. C. and Castelein, R. M., "Upright, prone, and supine spinal morphology and alignment in adolescent idiopathic scoliosis," Scoliosis Spinal Disord. **12**(6), 1–8 (2017).

Sophia Doerr | Curriculum Vitae

sophia@doerr.us

EDUCATION

M.S.E, Biomedical Engineering, Johns Hopkins University, 2020
Concentration: Imaging & Instrumentation

B.S., Biomedical Engineering, Applied Mathematics & Statistics, Johns Hopkins University, 2020
Concentrations: Computational Biology, Statistics & Statistical Learning
Minor: Computational Medicine

RESEARCH

Research Assistant, Johns Hopkins University ISTAR Lab, 2018-2020
Developed methods for automatic assessment of global spinal alignment from CT images and for automatic detection of spinal implants in radiographs

Senior Independent Design Project, Johns Hopkins University August 2016- May 2017
Prototyped and designed a cost-effective and reliable preterm labor screening device
Achieved reliability of 67% compared to standard of care using measure of cervical consistency index

Electrical and Computer Engineering (ECE) Lab Assistant: July 2016- May 2017
Prepared new microprocessor oscilloscope to use for Fall 2016 ECE lab course
Managed electrical components, test equipment, and program microprocessors in Python and Matlab

INDUSTRY

Cloud-based Software Consultant, Applied Physics Lab January 2018- August 2018
Conducted UX interface testing with manual and automatically-scripted processes
Developed and employed AWS data analytic algorithms for experiment with thousands of participants

Biomedical Engineering Summer Intern, Medtronic (Littleton, MA) May 2017- August 2017

Prepared model-based training imaging pipeline in C# for network training data
Trained Network with Dice score of 0.86

PUBLICATIONS AND RESEARCH AWARDS

Doerr SA, De Silva T, Vijayan R, Han R, Uneri A, Ketcha MD, Zhang X, Khanna N, Westbroek E, Jiang B, Zygourakis C, Aygun N, Theodore N, Siewerdsen JH "Automatic analysis of global spinal alignment from simple annotation of vertebral bodies". Journal of Medical Imaging. (Accepted 2020).

S. A. Doerr, A. Uneri, Y. Huang, C. K. Jones, X. Zhang, M. D. Ketcha, P. A. Helm, J. H. Siewerdsen, "Data-driven detection and registration of spine surgery instrumentation in intraoperative images," Proc. SPIE 11315, Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling, 113152P (2020)

Data-driven detection and registration of spine surgery instrumentation in intraoperative images
Best Poster Award for Image-Guided Procedures, Robotic Interventions, and Modeling
SPIE Medical Imaging 2020, Houston, TX